Economic decisions and judgments on road safety and health: a psychological approach

José Antonio Robles Zurita

Thesis Advisors:

PhD José Luis Pinto Prades

PhD José María Abellán Perpiñán

Departamento de Economía, Métodos Cuantitativos e Historia Económica

Universidad Pablo de Olavide

Sevilla, Abril 2014

Agradecimientos

Quiero agradecer sinceramente a José Luis Pinto Prades por su dirección y apoyo en la realización de esta tesis y además por permitirme tener la suerte de trabajar con él en investigaciones previas. También le estoy enormemente agradecido a José Ignacio García Pérez por todo lo aprendido de él, sobre todo en econometría, y por su apoyo desde que tuve la oportunidad de conocerlo. Este trabajo también se debe a la dirección de José María Abellán Perpiñán y al equipo de Economía de la Salud de la Universidad de Murcia incluyendo a Jorge Eduardo Martínez, Fernando Ignacio Sánchez e Ildefonso Méndez. Concretamente este equipo realizó la encuesta en que se basan los dos primeros capítulos de esta tesis y a ellos les debo su ayuda para distintos análisis realizados.

También han contribuido directa o indirectamente los profesores y personal del departamento de Economía, Métodos Cuantitativos e Historia Económica de la Universidad Pablo de Olavide. En especial agradezco a Marisa Hidalgo, Antonio Villar, Juande Moreno, Dunia López, Paula González, Laura Riesgo, Manuel Hidalgo, Jesús Rodríguez y Benedetto Molinari. También agradezco a todos los profesores de los que he aprendido durante la realización del Máster de Economía de esta misma universidad. Puedo citar a Carlos Usabiaga, Diego Romero, Alfonso Sánchez, Francisco J. André, Francisco Gómez, Santiago Budría, Arantza Gorostiaga, Javier Pérez, Carmen Herrero, Pablo Amorós, Bernardo Moreno, Socorro Puy y José Luis Torres. Mis agradecimientos van para María Eugenia Angulo por su ayuda en cualquier tema administrativo.

También agradezco los comentarios y recomendaciones respecto al primer capítulo de esta tesis de Michael Jones-Lee de la Universidad de Newcastle en Reino Unido. Además, mi gratitud va para Neil Stewart de la Universidad de Warwick por la lectura y comentarios del tercer capítulo. Estoy muy agradecido al personal investigador y administrativo del Yunus Centre for Social Business and Health de la Glasgow Caledonian University que me acogieron durante una enriquecedora estancia investigadora. Al personal del departamento de Fundamentos del Análisis Económico de la Universidad de Alicante le estoy igualmente agradecido.

Muchas gracias también a mis compañeros de doctorado con los que he compartido “alegrías y angustias” derivadas de la tesis, y de nuestras investigaciones en general, durante las comidas en la cafetería. Quisiera agradecer a Daniel Oto, Alejandro García, Manuela Prieto, David Troncoso y María Victoria Avilés.

Por último, mil gracias a Vero, mi esposa, por su apoyo durante todo este tiempo. Y gracias a mi familia, a mis padres, suegros y hermanos.

Acknowledgements

I want to sincerely thank José Luis Pinto Prades for his guidance and support in completing this thesis and also for letting me have the pleasure of working with him in previous research. I am also extremely grateful to José Ignacio García Pérez for everything learned from him, especially in econometrics, and for their support since I met him for the first time. I am indebted to José María Abellán Perpiñán and other members of the Health Economics group from University of Murcia including Jorge Eduardo Martínez , Fernando Ignacio Sánchez and Ildefonso Méndez. Specifically this group conducted the survey analysed in the first two chapters of this thesis.

Also teachers and staff at the department of Economics, Quantitative Methods and Economic History at Pablo de Olavide University contributed directly or indirectly. In particular I thank Marisa Hidalgo, Antonio Villar, Juande Moreno, Dunia López, Paula González, Laura Riesgo, Manuel Hidalgo, Jesús Rodríguez and Benedetto Molinari. I also thank all the teachers of the Master in Economics at the same university. I may mention Carlos Usabiaga, Diego Romero, Alfonso Sánchez, Francisco J. André, Francisco Gómez, Santiago Budría, Arantza Gorostiaga, Javier Pérez, Carmen Herrero, Pablo Amorós, Bernardo Moreno, Socorro Puy, and José Luis Torres. My thanks go to María Eugenia Angulo for their help in any administrative issue.

I also appreciate the comments and suggestions, regarding the first chapter of this thesis, by Michael Jones-Lee from Newcastle University in the UK. Also, my gratitude goes to Neil Stewart from the University of Warwick for comments on the third chapter. I am very grateful to researchers and administrative staff from the Yunus Centre for Social Business and Health at Glasgow Caledonian University who welcomed me during a research stay. I am also grateful to the staff from the Department of Economics at University of Alicante.

Thanks to my PhD colleagues: Daniel Oto, Alejandro García, Manuela Prieto, David Troncoso and María Victoria Avilés.

Finally, many thanks to Vero, my wife, for her support during all this time. And thanks to my family, especially my parents and parents in-law.

Resumen (en español)

Una tarea primordial de los economistas es el estudio de cómo asignar recursos limitados de manera eficiente. Esto es particularmente importante para la toma de decisiones públicas, dada la existencia de un presupuesto limitado que tiene que ser asignado a las alternativas de gasto público. Durante los últimos años las administraciones públicas han tenido que enfrentarse a dificultades adicionales para mantener el suministro de bienes y servicios públicos debido a la crisis económica que comenzó en 2008. Consecuentemente los recursos públicos se deben asignar a las opciones más beneficiosas siguiendo criterios racionales.

Hay una opinión generalizada entre los investigadores y profesionales de que el valor de los beneficios de los programas públicos debe basarse en las preferencias de la población en general a fin de lograr la eficiencia. Dos enfoque principales existen: la Preferencia Revelada (PR) y la Preferencias Declaradas (PD). Los métodos basados en la PR se basan en el comportamiento microeconómico real de los destinatarios de los servicios públicos en el mercado, mientras que la PD se basa en las preferencias de los sujetos declaradas en decisiones y respuestas a un cuestionario. En esta tesis trato de contribuir a esta cuestión mediante el estudio de algunos métodos de la PD destinados a la obtención de preferencias en el contexto de la salud y la seguridad vial. Este es un punto crucial, dado que si queremos tomar decisiones basadas en los resultados obtenidos por estos métodos necesitamos comprender sus pros y sus contras y, en definitiva, sus propiedades.

En los dos primeros capítulos dos de los métodos principales que se utilizan en Economía de la Salud y la evaluación de programas de seguridad son objeto de estudio: la Valoración Contingente (VC) y una versión modificada de la Lotería Estándar (LEM). En el tercer capítulo se analiza otro de los métodos más importantes en Economía de la Salud: el *Time Trade Off* (TTO). El principal interés de esta tesis es el estudio de la consistencia de las valoraciones de la salud por diferentes métodos. En este sentido trato de responder a preguntas como: ¿Es la valoración de la seguridad vial diferente para la VC y la LE? ¿Cambian estas valoraciones al variar algunas de las características de los procedimientos utilizados? Cualquier inconsistencia se interpreta como un desajuste del modelo económico estándar de las preferencias individuales. Como consecuencia de ello propongo algunos modelos de la Psicología y la Economía del Comportamiento que pueden explicar el desajuste entre los distintos métodos.

Con respecto al Capítulo 1 analizo la VC y la LEM. Los resultados indican que las valoraciones son muy distintas. En particular, los valores estimados a través de la LEM son muy inferiores a los valores derivados de la VC. Este es un resultado consistente con estudios previos. En la búsqueda de una justificación de este resultado considero el caso en que los individuos se comportan de acuerdo a *Prospect Theory* en lugar de seguir los supuestos de la Teoría de la Utilidad Esperada. Tras derivar las valoraciones teóricas encuentro una relación diferente entre los métodos utilizados. Los resultados indican que el uso de *Prospect Theory* como la base para el cálculo de las valoraciones por la LEM hace acercar este método a la VC. No obstante, persisten diferencias entre la LEM y la VC, y por lo tanto rechazamos a *Prospect Theory* como la única explicación. Por otro lado, y de manera sorprendente, se obtiene un mejor ajuste de los valores de la VC por las respuestas a una Escala Visual Analógica (EVA). Un análisis de consistencia pone de manifiesto que la VC es más insensible a la gravedad de los estados de

salud que se valoran. Al mismo tiempo, las personas cometen más errores en el método de la LEM.

En el segundo capítulo, examino el efecto del modo de evaluación (ME) a través del análisis de la Evaluación Conjunta (EC) y Evaluación Separada (ES) para ambos procedimientos, la VC y la LEM. La sensibilidad de las valoraciones de los estados de salud es más alta en EC tanto para la VC como para la LEM. Sin embargo, la insensibilidad encontrada en ES es mucho más extrema para la VC. En el mismo capítulo 2 comparo también diferentes grupos de EC para analizar los efectos de contexto. Los resultados indican que los efectos de contexto afectan más a las respuestas de la VC que las valoraciones de la LEM. En la última parte del capítulo analizo la valoración de la seguridad vial en función del escenario de renta en que se encuentran los individuos: un escenario de las pérdidas (con ingresos corrientes inferiores a los ingresos normales o medios); un escenario de ganancias (con ingresos corrientes superiores a los ingresos normales); y un escenario neutral (con ingresos corriente iguales a los ingresos normales). Los resultados indican que aquéllas personas en un marco de ganancias indican una mayor satisfacción en la vida que las de un marco neutral o de pérdidas. Es más, la disposición a pagar para evitar el riesgo de una muerte o lesión en carretera, derivadas de las repuestas a la VC, es mayor para los individuos en un escenario de ganancias. Este resultado se puede explicar por una función de utilidad que depende de los ingresos de referencia (renta permanente o renta promedio). Por el contrario no se encuentran efectos para el caso de la LEM.

Finalmente en el capítulo 3 se analizan los datos de un experimento basado en una encuesta *online* para estudiar si el método TTO está afectado por los efectos de contexto predichos por el modelo de la *Rage Frequency Theory* (RFT). El atractivo de esta teoría es que podemos calcular el “valor real subyacente” para los estados de salud objeto de valoración. Los resultados encontrados en esta tesis sugieren que RFT predice correctamente los efectos del contexto. Sin embargo, la aplicación de RFT a los datos no deriva en valoraciones "libres de contexto".

En conclusión, esta tesis estudia sesgos e inconsistencias entre los procedimientos de obtención de valoraciones en salud. Estos sesgos no son tenidos en cuenta por el modelo estándar en Economía, sin embargo su existencia es un desafío para el análisis económico. Se deberían establecer criterios normativos para la elección de los procedimientos más adecuados.

Introduction

A primary task of economists is the study of how to allocate limited resources efficiently. This is particularly important for public decision making given the existence of a limited budget that has to be assigned to public expenditure alternatives. During the last years, governments have had to face additional difficulties to maintain the supply of public goods and services due to the economic crisis begun in 2008. Consequently public resources have to be assigned to those more beneficial choices following rational criteria. For this purpose we have to perform Cost-Effectiveness (CE), Cost-Utility (CU), and Cost-Benefit (CB) analysis. CE requires assessing the costs and the beneficial outcomes of public interventions. CU requires estimate the cost and the utilities derived from public programs. Eventually CB transforms any cost and favourable outcome to monetary values. CE and CU allow us to discriminate between different programs and choose the most profitable alternative. In addition CB can make a judgment on whether a specific public program should be undertaken.

There is a widespread opinion among researchers and professionals in the field that the value of the benefits of public programs should be based on preferences of the general population in order to achieve efficiency. Two main approaches to this valuation task have been given in the economic literature, namely Revealed Preference (RP) and Stated Preference (SP). Methods based on RP are based on actual microeconomic behaviour of recipients of public services in the market while SP relies on subjects' preferences elicited by a survey questionnaire. In this thesis I try to contribute to the matter by studying state preference methods aimed at eliciting preferences in the context of health and road safety domains. This is a crucial point given that if we want to make decisions based on results given by these SP methods we need to understand their pros and cons and eventually their properties.

In Chapter 1 and 2 two major methods used in health economics and road safety programs assessment are studied: Contingent Valuation (CV) and a modified version of the Standard Gamble (SG). In the third Chapter I analyse the other important method in health economics: the Time Trade Off (TTO). The main interest for the present thesis is to study the (in)consistency of valuation of health between and within different methods. In this sense I try answer to questions like: Is road safety valuation different by CV and SG? Is road safety or health valuation changed when varying some features of the elicitation procedures? Any inconsistency is interpreted as a fail of the standard economic model of individual preferences. As a consequence I study whether some models from psychology and behavioural economics can account for mismatch between different valuations. Given that several methods give place to distinct health utilities and valuation of road safety I consider whether we can rely on those psychological models to estimate what is called "true underlying values".

Road safety policy is a major concern for the global economy since traffic accidents lead to huge losses in human and financial costs. In 2010 about 1.24 million people were killed and between 20 and 50 million suffered a non-fatal injury as a result of traffic crashes (World Health Organization, 2013). This is a harder problem for middle income countries due to their rapid motorizing. However in developed countries the rate of road fatalities is about 12 in 100,000 people (see Cubí-Mollá and Herrero, 2008). Specifically in Spain 117.793 people suffered damage as a consequence of a road accident in 2012 and 1.903 of them died as a result (see *Dirección General de Tráfico,* 2013). Despite the fact that these figures have decreased in the last years road accidents continue to be a major cause for health losses.

In this sense, Chapter 1 and 2 present two separate analyses of data from a survey study founded by the DGT (Spanish Road traffic Directorate General) that was interested in the estimation of the value of preventing Non-Fatal Road Injuries (NFRI) to apply this value to investment programs assessment in the road safety domain. The number of participants was 2,016 and the issues that can be studied are various. The study of two value elicitation procedures is the focus of Chapter 1, namely: Modified Standard Gamble (MSG) and Contingent Valuation (CV). Chapter 2 analyses the effects on the value of preventing road accidents of: a) the evaluation mode, either Separate or Joint; b) the context effects given by contextual injuries included in Joint evaluation, and; c) income frame derived from location of current income situation of respondents with respect to their permanent or reference income.

With respect to Chapter 1 Contingent Valuation (CV) and the so called Modified Standard Gamble (MSG) are supposed to be valid for the estimation of the *relative value* of preventing a NFRI with respect to the value of preventing a traffic fatality, also known as the value of the statistical life. CV involves asking respondents about the amount of money they are willing to pay (WTP) for reducing the risk of a NFRI in one unit ($\boldsymbol{m_I}$), on the one hand, and for reducing the risk of death ($\boldsymbol{m_D}$), on the other hand. MSG estimates the relative value $\left(\frac{\boldsymbol{m_I}}{\boldsymbol{m_D}}\right)$ in a more indirect way by asking individuals the amount of risk of death (π^*) they are willing to accept in order to avoid a situation in which they suffer a NFRI with a small probability of death (θ). Under expected utility (EUT) the MSG *relative value* should be computed as

$$\frac{\boldsymbol{m_I}}{\boldsymbol{m_D}} = \frac{\pi^* - \theta}{1 - \theta}. \tag{1}$$

However I find that MSG relative values computed as (1) vary substantially from relative values computed directly from CV responses. In particular MSG values are much lower than CV ones. This is a result consistent with previous studies (Jones-Lee et al., 1995). In the quest for a justification of the CV-MSG disagreement I consider the case in which individuals are not expected utility maximizers but rather they weight probabilities and have loss aversion as in Prospect Theory (see CPT by Tversky and Kahneman, 1992; and early version of PT also by Kahneman and Tversky, 1979). For that purpose I derive theoretical relative values and MSG relative values of preventing a non-fatal injury under a CPT framework and a different relation between CV and MSG is shown. In the realization of this task I apply the evaluation of lotteries as previously done in Bleichrodt et al. (2001 and 2007) in empirical implementations of Prospect Theory for health states valuation. In these implementations it is assumed that the Reference Point is certain, for example one health outcome or health state, according to which the rest of the outcomes or states are considered as gains or losses. I improve this analysis by allowing the reference to be uncertain as in Third Generation Prospect Theory (PT^3) by Schmidt et al (2008). Since in the survey design respondents have to choose between two risky lotteries, in both CV and MSG, they could take one of them as the reference.

Results indicate that the use of Prospect Theory as the basis for computing MSG relative values reduces the gap between CV and MSG. Eventually, even though PT gains descriptive power, differences still remain between CV and MSG, and thus we reject PT to be an explanation. Surprisingly, I obtain that a much better fit of CV relative values is obtained if we use responses to a Visual Analogue Scale and consider them as “utilities”. So CV seems to entail an evaluation process much more similar to a rating scale than to a choice-based elicitation procedure like MSG. Also a consistency analysis shows up that CV is more insensitive to severity of injuries (people are much willing to pay the same for avoiding a risk of a mild health state as for avoiding a severe injury) than MSG responses. At the same time individuals commit more errors in the MSG method because a higher percentage of them do an evaluation that is just opposite to their preference order on injuries previously shown in a ranking exercise.

In the second Chapter I examine the effect of the Evaluation Mode (EM) through the analysis of Joint Evaluation (JE) and Separate Evaluation (SE) for both elicitation procedures MSG and CV. These two EMs has been shown to differ in several respects. For example, preference reversal occurs between the two modes as Hsee (1996) reported in an experiment in which subjects had to assign a salary to candidates for a job. Also insensitive valuation responses are more likely to occur in SE as proposed by General Evaluability Theory (see GET of Hsee and Zhang, 2010). I find that the value sensitivity is higher in JE for both CV and MSG. However, value insensitivity encountered in SE is much more extreme for CV given that some injuries with varying severity are not evaluated significantly different. While for MSG even in SE high value sensitivity remains because injuries are evaluated significantly different. Therefore valuation of NFRIs seems to be less affected by the specific EM which is a factor to be considered for a suitable State Preference elicitation procedure. In addition, I find that a systematic effect of the EM is that those mild (serious) NFRIs are evaluated as more (less) severe in SE than in JE. This last result suggests that SE evaluation of milder (more severe) injuries should be downward-corrected (upward-corrected) to encounter JE.

In the same Chapter 2 I also compare different JE groups to test for context effects. I test whether these effects are consistent with contrast effects predicted by Parducci’s range-frequency model (see Parducci, 1965). I also consider the possibility of anchoring context effects consisting on the empirical finding that judgmental responses given by individuals are affected by initially presented values (Lichtenstein and Slovic, 1971). Results indicate that context effects between different JE groups do affect more to CV responses than MSG valuations. So not only MSG seems to be less affected by JE-SE modes but within JE it is again more consistent between different contexts than CV. Neither Parducci’s range frequency theory nor anchoring seem to be a consistent explanation for valuation shifts across JE groups.

In the last part of the Chapter I analyse valuation of respondents that are framed in different income situations: a frame of losses; a frame of gains; and a neutral frame (no gain, no losses). I find that those in a gain frame (with current income higher than normal income) report a higher life satisfaction than those in a neutral frame (with current income equal to normal income) and than those in a losses frame (with current income lower than normal income). I propose that a plausible explanation is that respondent used their permanent income as a reference point with respect to which they evaluate their current income. This result is consistent with the fact that happiness depends on relative income rather than absolute income (see Easterlin 1974 and 1995). Most importantly the rest of the analysis is aimed at exploring if this framing effect on reported happiness is also relevant for decisions that should be based on utility. I find that WTP for avoiding a risk of a road fatality or road injury is higher for those in a gain frame with respect to those respondents in a neutral and loss scenario. I show that this result can be explained by a general reference dependent utility function of income with the typical properties of loss aversion and diminishing sensitivity (Kahneman and Tversky, 1979; and Tversky and Kahneman, 1992). On the contrary I do not find income frame effects in MSG responses for the majority of the injuries.

Eventually in Chapter 3 I analyse data from an experiment based on an online survey to study whether TTO is hindered by range frequency theory effects. Specifically I try to address questions like ¿Are health states utilities context-dependent? ¿What theoretical model is a good account of this phenomenon? And if utility elicitation does vary with context ¿Is it possible to correct for context effects in order to elicit the so called “context-free” health utilities? I apply range frequency theory (RFT; Parducci, 1965) to manipulate the contextual health profiles and change the so called frequency value across contexts. The attractiveness of this theory is that we can compute the underlying true value for the target stimulus; in our case the health state. In the health domain previous studies have found context effects related to manipulation of the frequency value (Bleichrodt and Johannesson, 1997; Robinson et al., 2001). However, in general there is a lack of evidence making clear if utility elicitation techniques used in Health Economics, like SG or Time Trade Off (TTO), are subjected to RFT effects. The results here suggest that RFT predicts properly context effects while other theory of contrast effects like Adaptation Level Theory (Helson, 1964) fails to predict some specific effects. However the fitting of RFT to data does not derive in invariant “context-free” utilities. Specifically, these context-free utilities are the same for groups with the same number of contextual health profiles but they differ between groups with different number of contextual health profiles.

In conclusion in this thesis I study biases and inconsistencies between elicitation procedures (MSG vs. CV), evaluation modes (JE vs. SE), income frames (frames of loses, gain or neutral frame), and context (manipulation of distribution of contextual health profiles). These biases are not considered by the standard economic analysis however their existence is a challenge for economic valuation since monetary values or elicited utilities that should be the basis for public decision making (e.g. health programs) depend on specific features of the valuation process. Normative criteria should be set to be the basis for choosing the appropriate elicitation procedure, evaluation mode or contextual features. For example, MSG seems to give more sensitive responses and be less affected by the evaluation mode and contexts which are positive attributes for an elicitation technique. On the other hand, there may not exits the so called underlying preferences or context free valuation but rather every valuation need a context or other features included in the valuation process. In this sense, the task for the economic analyst is to find which is the appropriate context or feature that should be present in the valuation.

References

1. Bleichrodt, H., Abellan-Perpinan, J. M., Pinto-Prades, J. L., and Mendez-Martinez, I. (2007). Resolving inconsistencies in utility measurement under risk: Tests of generalizations of expected utility. *Management Science, 53*(3), 469-482.
2. Bleichrodt, H., and Johannesson, M. (1997). An experimental test of a theoretical foundation for rating-scale valuations. *Medical Decision Making*, *17*(2), 208-216.
3. Bleichrodt, H., Pinto, J. L., and Wakker, P. P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science*, *47*(11), 1498-1514.
4. Cubí-Mollá, P., and Herrero, C. (2008). *Evaluación de riesgos y del impacto de los accidentes de tráfico sobre la salud de la población española, 1964-2004*. Fundación BBVA.
5. Dirección General de Tráfico (2013). *Anuario Estadístico de Accidentes, 2012*. Servicio de Estadística. Observatorio Nacional de Seguridad Vial.
6. Easterlin, R.A. (1995). Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior and Organization*, 27, 35–47.
7. Easterlin, R.A. (1974). Does economic growth improve the human lot? In P.A. David and M.W. Reder (Eds.), Nations and households in economic growth: Essays in honor of Moses Abramovitz (pp. 89–125). New York: Academic Press.
8. Helson, H. (1964). Current trends and issues in adaptation-level theory. *American Psychologist*, *19*(1), 26.
9. Hsee, C. K (1996). The evaluability hypothesis: An explanation of preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 46*, 247-257.
10. Hsee, C. K., and Zhang, J. (2010). General Evaluability Theory. *Perspectives on Psychological Science, 5*(4), 343-355.
11. Jones-Lee, M., Loomes, G., and Philips, P. (1995). Valuing the prevention of nonfatal road injuries - contingent valuation vs. standard gambles. *Oxford Economic Papers-New Series, 47*(4), 676.
12. Kahneman, D., and Tversky, A. (1979). Prospect theory - analysis of decision under risk. *Econometrica, 47*(2), 263-291.
13. Lichtenstein, S., and Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89*(1), 46-55.
14. Parducci, A. (1965). Category judgement: a Range-Frequency model. *Psychological Review*, 72(6), 407-418.
15. Robinson, A., Loomes, G., and Jones-Lee, M. (2001). Visual analogue scales, standard gambles, and relative risk aversion. *Medical Decision Making*, *21*(1), 17-27.
16. Schmidt, U., Starmer, C., and Sugden, R. (2008). Third-generation Prospect Theory. *Journal of Risk and Uncertainty, 36*(3), 203-223.
17. Tversky, A., and Kahneman, D. (1992). Advances in prospect-theory - cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297-323.

18. World Health Organization (2013). *WHO global status report on road safety 2013: supporting a decade of action*. World Health Organization.

INDEX

Chapter 1. Valuation of preventing road injuries: a Cumulative and Third Generation Prospect Theory approach

1. Introduction

The Spanish Road traffic Directorate General (*Dirección General de Tráfico*, DGT) is interested in eliciting the value of preventing non-fatal risks to apply this value to assessment of road safety investment programs. As a consequence, a survey study founded by the DGT was carried out for this commitment. This work is aimed at describing how respondents, Spanish population included in that survey, value the prevention of Non-Fatal Road Injuries (NFRIs) through two elicitation procedures well known in the literature concerned, namely: Contingent Valuation (CV) and the so called Modified Standard Gamble (MSG). The two procedures elicit the *relative value* of preventing a NFRI with respect to the value of preventing a traffic fatality (also known as the value of the statistical life). CV is a major procedure used for stated preferences elicitation when no market prices are available or these are no reliable. Particularly, CV has been widely used in other disciplines such as environmental assessment. In Arrow et al. (1993) some aspects of the CV implementation are illustrated and guidelines are suggested. Also Venkatachalam (2004) presents a more recent review of the method. The Standard Gamble (SG) is one of the most accepted utility elicitation method in Health Economics. In this method, a decision maker decides between a risky situation and risk-less option. The MSG used here is a variant where the decision is made between two risky lotteries (see Oliver, 2005, for an implementation and consistency test of the method).

In the work presented here we make use of these two methods in the traffic safety context as previously done in the literature. Specifically, we follow closely the methodology approach adopted by Jones-Lee et al. (1995) in a study carried out in the United Kingdom in which 414 respondents were interviewed with Contingent Valuation questions and other 409 did a Standard Gamble (SG) questionnaire.[1] The structure of the survey allowed between sample comparisons of both methods. They also carried out a "follow-up" survey with 101 individuals that had participated in one of the two samples, CV or SG, in order to value NFRI prevention with the alternative method, SG or CV respectively, allowing within sample comparisons for this subgroup. The results indicated that both elicitation procedures did not relate in the expected manner, and doubts about the theoretical assumptions for the relation between the methods, and about the procedures themselves, emerged. In fact, much higher relative values were computed by CV, being between about ten times, for the least severe injury denoted by W, and three times, for the severest injury denoted by R, higher than SG ones.

The theoretical relation between CV and SG (or MSG) responses is shown by Jones-Lee (1989) and is mainly based on the assumption that subjects' responses follow the principles of Expected Utility Theory (EUT). In the quest for a justification of the CV-MSG mismatch we consider the case in which individuals are not expected utility maximizers but rather weight probabilities and have loss aversion as in Prospect Theory (see PT by Kahneman and Tversky 1979). Specifically, we introduce in the analysis Cumulative Prospect Theory (CPT, Tversky

[1] This citation shall be referred to as the *British study* or *Jones-Lee study*.

and Kahneman, 1992) and Third Generation Prospect Theory (PT^3, Schmidt et al. 2008). We derive theoretical relative values of preventing a non-fatal injury under a CPT framework and a different theoretical relation between MSG and CV responses is found as a consequence. In the realization of this task we apply the evaluation of lotteries as previously done by Bleichrodt et al. (2001 and 2007) in empirical implementations of CPT for health states valuation. In these implementations it is assumed that the reference point is a certain health state (e.g. one of the health states included in a risky treatment). Accordingly the rest of the outcomes or health states are considered as gains (losses) if they are better (worse) than the reference. The previous findings suggest that PT is a descriptive preference theory more appropriate than EUT. On the one hand, Bleichrodt et al. (2001) were able to explain disparities between utilities elicited by three methods: *Probability equivalent, Certainty equivalent* and *Tradeoff*. In addition, Bleichrodt et al. (2007) found that utilities elicited by five methods (*Probability equivalent, Certainty equivalent¸ Value equivalent*, *Probability lottery equivalent* and *Value lottery equivalent*) were closer with a PT modelization of responses. We enhance our analysis by allowing the reference to be uncertain as in PT^3 by Schmidt et al. (2008). They show that PT^3 is valid to explain some stylized facts as preference reversal (see Lichtenstein and Slovic, 1971, for preference reversal with monetary values; and Stalmeier et al., 1997 with health outcomes) and inconsistency between Willingness To Pay (WTP) and Willingness To Accept (WTA) when a person is endowed with a risky lottery. Since in our survey respondents have to choose between two risky lotteries, in both CV and MSG, they could take one of them as the reference. Thus we explore whether this can be a good explanation for the inconsistency between the relative values of a NFRI estimated by the two methods.

The present article discusses the results of a survey that have certain differences with the *British Study* that may improve the understanding of the valuation of preventing NFRIs by society. First, the sample size more than doubles the UK sample size with a total of 2,016 respondents and they expressed preferences from both CV and MSG, which allows within sample comparisons of the two methods. Secondly, as just noted above, modified form of the SG was carried out rather than its classical type, in order to avoid the "certainty effect". This effect is shown when individuals are very reluctant to risk just a little their lives to avoid the certainty of an injury leading to a systematic response in some individuals who give the same answer without risking anything for a series of health states with markedly severity differences. With MSG this can be mitigated because the respondent does not have the choice of a riskless alternative, but two options that have some risk, and thus is more willing to risk something and discriminate more between health states that differ in severity. In third place, health states that represent possible injuries after a road accident are all valued directly in contrast to what happened in the UK study where some health states values were estimated by extrapolation from other injuries valuations. The fact that individuals face a wide range of health conditions makes it possible a comparative analysis of the two methods, CV and MSG, for conditions of varying severity. This aspect of the survey is crucial because the results we obtain indicate that percentage of consistent responses to the CV and MSG differs significantly with the severity of health states being valued, especially for those health states valued as worse than death by a large percentage of individuals.

We find that in case individuals make decisions as in PT (CPT or PT^3) then probability weighing and loss aversion have a role in the interpretation of MSG responses. However, the same theoretical relative value is found for the two theories of decision under risk, PT and EUT. Our results indicate much larger relative values estimated by CV as in Jones-Lee study. When using PT parameters previously estimated by Tversky and Kahneman (1992) MSG relative values do not make much better to fit CV estimations, although this theory gains descriptive power with respect to EUT. Neither the consideration of uncertain reference points explains the gap between the two valuation methods. When we estimate parameters that best fit our data MSG relative values are similar to CV ratios, but still statistically significantly different. Meanwhile values elicited by a Visual Analogue Scale (VAS), if interpreted as NFRIs utilities, and used for the computation of VAS relative values do a better fit of CV estimations than MSG. This result is somewhat surprising given that no clear theoretical connection between VAS and CV is considered in the literature.

We perform an analysis to explore to what extent underlying preferences in MSG and CV responses are consistent with a ranking task previously done. Consistency analysis of CV highlights that a major problem is the insensitivity of responses. Many respondents are willing to pay the same for the prevention of NFRIs even when those injuries vary considerably in severity both objectively, given the distance in consequences for health of these injuries, and subjectively, as shown in their ranking exercise. As a consequence individuals show a high degree of *embedding effect* which in turn seems to be an explanation for such higher CV relative values. For the MSG responses appear to be more sensitive to the variability in injury severity however they were more inconsistent than CV ones, this is a higher proportion of respondents showing preferences that are just opposite to those shown previously in a ranking exercise. Eventually, since MSG used in this study do not allow valuation of NFRIs worse than death, huge inconsistencies are found for the severest injuries suggesting that CV relative values are more reliable in those cases. In future surveys an adapted version of the MSG should be considered for the valuation of this worse than death health conditions.

Eventually, econometric analysis allows us to predict some consistency patterns followed by subjects according to their socio-demographic characteristics. For example, household income appears to be negatively related to embedding and thus suggesting a kind of budget constraint issue in responses to CV. Also inconsistency in MSG has been more likely for the group of respondents that do not understand probabilities (they fail in a previous testing question). These results indicate that it could be reasonable to provide a CV or MSG questionnaire to different subjects depending on their characteristics in order to estimate more accurate valuations. Also VAS score absolute differences between two NFRIs is a very good predictor of a consistent valuation in both CV and MSG so that an interval scale property is found for this method.

This work is organized as follows. In the next section, survey details are exposed. Theoretical considerations for the estimation of the relative values of preventing a NFRI are included in sections 3 to 5. Then relative value estimations are shown in the sixth section and consistency analyses are developed in section 7. Eventually section 8 includes a discussion and 9 concludes.

2. The survey

2.1. Design

The survey is conducted during the first quarter of 2011 (January-March) through interviews taken place in the home of the respondents with the help of a laptop where all the questions are illustrated by a computer program. A set of 55 questions divided into four parts are presented to a Spanish nationally representative sample of 2,016 individuals. The first part of the questionnaire collects information about the use of road transport by respondents and some comprehension questions. Secondly, respondents rank the NFRIs they are going to value and place them in a Visual Analogue Scale (VAS) ranging from 0 to 100. The core of the questionnaire is comprised by the Modified Standard Gamble (MSG) and Contingent Valuation (CV) questions through which respondents value four Health States.[2] Eventually, the rest of the questionnaire is aimed at collecting socio-demographic information of respondents.

The eight NFRIs described in Figure 1 are used for valuation.[3] These are analogous to those injuries included in Jones-Lee et al. (1995). Each Health State (HS) presents a different level of seriousness in some attributes like time in hospital, the extent and duration of pain, degree and length of restrictions to leisure and work activities, degree of physical and mental ability, and capacity to meet basic physical needs. So NFRIs extend over a wide range from the mildest ones, like F or W, to the most serious, like N or L. Accordingly, it may be considered that there is an objective preference order as $F \succcurlyeq W \succcurlyeq X \succcurlyeq V \succcurlyeq S \succcurlyeq R \succcurlyeq N \succcurlyeq L$. This idea is supported by the fact that it can be assumed that people's preferences over the seriousness of health attributes are monotone and that the eight NFRIs are increasing in the severity of the attributes. Nonetheless, we allow respondents to show different order of preferences to study their consistency across different elicitation methods.

Eight different questionnaires are presented, one to each different group of respondents. In Table 1 it is shown the differences between these eight groups that are due to the specific NFRIs valued. All the groups respond MSG questions to value the prevention of four HSs. With respect to CV questions, groups 1-4 respond to the valuation of preventing one separate NFRI while groups 5-8 respond to the valuation of preventing four different injuries. In addition, all the individuals respond to a CV question to value the prevention of risk of death that makes the computation of the relative values of preventing the NFRIs possible.

Before valuation questions are carried out respondents are given information about the actual risk of traffic accident in Spain differentiating between fatal and non-fatal injuries so that they get familiar with the current state of road safety. In addition they were given information about different causes of death (e.g. different types of cancer, heart diseases, and diabetes among others) to compare with risk of a road fatality (in the yellow box) as in Figure 2.

[2] In what follows the damage after an accident, prevention of which is being valued, shall be appointed either as health profile, health condition, health state (HS) or NFRI.
[3] All the interviews are carried out in Spanish though in this thesis it will be shown the English translations of questions, injuries descriptions, or instructions.

<table>
<tr><th>F</th><th>W</th></tr>
<tr><td>Does not require hospitalization, the patient is treated in outpatient settings.
After Effects:
Mild to moderate pain for 1 week.
There are difficulties in work and leisure activities that gradually reduce.
After 3 or 4 months, full recovery without any sequelae.</td><td>In hospital:
1 week
Mild pain
After Effects:
Pain or discomfort for several weeks.
There are difficulties in work and leisure activities that gradually reduce.
After 3 or 4 months, full recovery without any sequelae.</td></tr>
<tr><th>X</th><th>V</th></tr>
<tr><td>In hospital:
2 weeks
Moderate pain
After Effects:
Pain gradually reduces.
There are difficulties in work and leisure activities that gradually reduce.
After 18 months, full recovery without any sequelae.</td><td>In hospital:
2 weeks
Moderate pain
After Effects:
moderate to severe pain for 1-4 weeks
Then, the pain gradually fades, but reappears when performing certain activities.
There exist permanent restrictions to work and leisure activities.</td></tr>
<tr><th>S</th><th>R</th></tr>
<tr><td>In hospital:
4 weeks
Moderate to severe pain
After Effects:
moderate to severe pain for 1-4 weeks
Then, the pain gradually fades, but reappears when performing certain activities.
There exist permanent restrictions to work and leisure activities.</td><td>In hospital:
More than 4 weeks, possibly several months
Moderate to severe pain
After Effects:
Lifelong chronic pain
There are major and permanent restrictions to work and leisure activities.
Possibly some prominent and permanent scars.</td></tr>
<tr><th>N</th><th>L</th></tr>
<tr><td>In hospital:
More than 4 weeks, possibly several months
Inability to use the legs and arms, possibly due to paralysis or amputation.
After Effects:
Confined to a wheelchair for the rest of life
Dependent on others for many physical needs such as dressing and toileting</td><td>In hospital:
More than 4 weeks, possibly several months
Head injuries that cause permanent brain damage
After Effects:
Mental and physical abilities greatly reduced for the rest of your life.
Dependent on others for many physical needs such as dressing and toileting</td></tr>
</table>

Figure 1. Non-Fatal Road Injuries (NFRIs) for valuation

Table 1. Sample size and Non-Fatal Road Injuries valued by group

		NFRIs	
Group	**Obs.**	**MSG**	**CV**
1	254	F, W, N, L	F
2	251	X, V, N, L	V
3	256	X, V, S, R	X
4	251	F, W, S, R	R
5	253	F, W, N, L	F, W, N, L
6	250	X, V, N, L	X, V, N, L
7	248	X, V, S, R	X, V, S, R
8	253	F, W, S, R	F, W, S, R
Total	**2,016**		

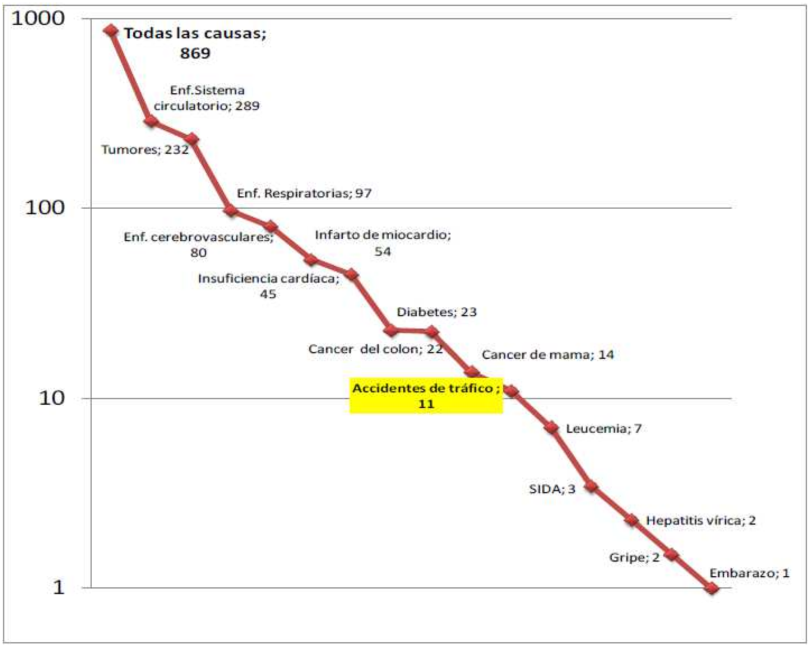


Figure 2. Causes of death in Spain. Number of fatalities in 100,000. Logarithmic scale

2.2. The ranking and VAS exercise

The ranking and VAS exercises are presented before the valuation by CV and MSG is carried out so that both: respondents become familiar with health states; and information about their preferences is obtained. On the one hand, the ranking lets us know the order of preferences over the four health states that each respondent has to value, on the other hand, the visual analogue scale provides us information about the strength of those preferences. In addition respondents value death state and their normal health ("health today") along with the NFRIs.

For the ease of the ranking task those interviewed are provided a set of six cards with the description of each of the situations to be considered (4 NFRIs plus normal health and death), arranged at random, and asked to rank the six cards on the table from worst to best. Subsequently, the order is transferred by the interviewer to the software application.

For the VAS, respondents place each of the six health conditions at a point along the line of the visual scale between 0 and 100. They are given instruction to reserve the extreme points, 0 and 100, to the worst and best health condition they could imagine, respectively. Also, they are given instructions to assign a VAS score to each health state such that the distance between those reflects the strength of their preferences (i.e. if W is worse than F, but R is much worse than W, then the distance between W and F has to be shorter than the distance between R and W). These requirements for individual responses are also presented in the EQ-5D Visual Analogue Scale (see Brooks et al. 2003) and are supposed to be necessary and sufficient for providing a measure of preferences with the properties of an interval scale as argued by Parkin and Devlin (2006) and earlier by Jones-Lee et al. (1994). The ranking exercise precedes the rating task which improves the quality of VAS responses, as it has been argued in the literature (see Greiner, 2003). In Figure 3 there is an example of the VAS exercise.

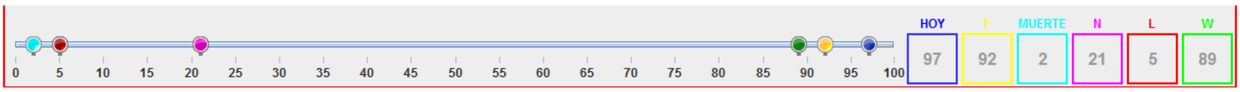


Figure 3. VAS exercise. Note: *"hoy"* and *"muerte"* stand for normal health and death

2.3. Sample characteristics

In Table 2 it is presented some of the sample characteristics. First we will describe the sample with respect to the general socio-demographic characteristics: gender, age, education, work status and income. For gender and age it is also shown the distribution of the Spanish population according to the 2011 Census of Population. In our sample there is slightly more females than males in line with corresponding population figures. All subjects are above 18 and are very similar to the Census population with respect to age as can be noticed. Almost half of respondents have lower secondary education, primary education or no education whereas the rest of the sample equally split between upper secondary education and tertiary education. For the work status 60% of the subjects belong to the active population and the majority are workers in the private sector followed by self-employed and workers in the public sector. Also 12% of the sample is unemployed. With respect to monthly household income the sample is approximately divided into two halves: those above or below €1,200.

We also collect information about other characteristics more specific to safety context. For example, previous to the valuation of safety risks, a “test question” is presented to subjects in order to know whether they understand probability concepts. The question is: “*Imagine that the probability of dying from a car accident is 1% (1 in 100 fatal accidents). In this situation, how many people would die for each group of 1,000?”*. We expect *10* to be the correct answer. The same question with a different probability is presented subsequently. The huge majority, 96%, answer correctly both questions, while the rest show difficulties in the understanding of risks. Most of the respondents drives a car and have suffered a car accident. Almost one third of the sample are smokers, most of them smoke less than 10 cigarettes a week. About 70.4% of interviewed subjects play gambling games at least with certain frequency. Slightly more than half of the sample practices sport. Eventually, subjects reported their subjective health status. Most of the subjects, 71.8%, report good or very good health. In the extreme, 13.2% report excellent health and 15% moderate or bad.

Table 2. Percentage distribution of sample characteristics

Variables	Sample (%)	Census (2011)	Variables	Sample (%)
Gender:			**Driver:**	
Male	48.7	49.3	Yes	61.8
Female	51.2	50.6	No	38.2
Age:			**Suffered accident:**	
18-29	17.7	16.1	Yes	72.7
30-39	20.3	20.2	No	27.3
40-49	20.7	19.4		
50-65	23.7	23.3	**Smoker:**	
>=66	17.4	20.9	Non smoker	66.8
			1-10 cig. a week	13
Education:			11-20 cig. a w.	12.6
No education, Prim. or Lower Sec.	47.8		>20 cig. a w.	7.6
Upper Secondary	27.9			
Tertiary	24.2		**Plays gambling**:	
			Yes	70.4
Employment Status:			No	29.6
Inactive (cons.)	40.1			
Worker: private sector	32.9		**Practices sports:**	
Worker: public sector	6.2		Yes	56.7
Self Employed	8.6		No	43.3
Unemployed	12.2			
			Self-reported Health:	
Household income (€):			Excellent	13.2
0 – 1200	49.2		Very Good	33.7
1200 – 1800	25.3		Good	38.1
> 1800	25.5		Moderate	12.9
			Bad	2.1
Understand risk:				
Yes	96			
No	4			

3. Theoretical Relative Value of preventing a Non-Fatal Road Injury under Expected Utility Theory

In this study, the estimation approach is the *relative value* as in the *British Study*.[4] It is based on the computation of a ratio for each NFRI which is interpreted as the value of preventing the risk of non-fatal accidents relative to the value of preventing the risk of a fatality in a traffic accident. For example, if the value of preventing a particular risk of non-fatal accident is m_I and the value of preventing the same amount of a fatality risk is m_D, also known as value of a statistical life, then the relative value will be $\frac{m_I}{m_D}$. Once estimated the relative value, multiplication of that ratio by the value of a statistical life would be enough to get the value of preventing NFRIs to be taken into account for Public Spending decisions.[5] In this section we are interested in the theoretical basis for individual relative values of preventing an injury under the Expected Utility hypothesis.

In the survey two methods are considered for estimating the relative value: MSG and CV. Both are based on the theoretical framework developed by Jones-Lee (1989) and Carthy et al. (1999). In this framework it is assumed that the individual is an expected utility maximizer and faces a lottery in which a possible state S is defined by two outcomes, health profile (hp) and wealth (w), $S = \{hp, w\}$. Three possible states are considered, each one with a different hp: the subject could die in a road accident, denoted by $hp = D$; he could suffer a NFRI so that $hp = I$; or he enjoys his normal health with $hp = U$.

Assume that individuals have a probability $\overline{q}$ of having state $\{I, \overline{w}\}$ otherwise having their current state $\{U, \overline{w}\}$, then this is a lottery that can be expressed as $A = [\overline{q}, \{I, \overline{w}\}; 1 - \overline{q}, \{U, \overline{w}\}]$ and its expected utility, $\overline{EU}$, is

$$\overline{EU} = (1 - \overline{q})U(\overline{w}) + \overline{q}I(\overline{w}) \tag{1}$$

where $U(\overline{w})$ and $I(\overline{w})$ are the utility of initial wealth, $\overline{w}$, conditional on having normal health and an injury, respectively.

Suppose now that this individual is given the chance to change the initial probabilities of having a non-fatal injury from $\overline{q}$ to q. This individual would be willing to change his wealth to w as long as his expected utility does not vary.[6] So the new situation satisfies:

$$\overline{EU} = (1 - q)U(w) + qI(w). \tag{2}$$

Now we want to compute the Marginal Rate of Substitution (MRS) of wealth for risk of injury given by the indifference curve at the situation where the risk of non-fatal accident is $\overline{q}$. In other words, we calculate the derivative of wealth with respect to the risk of NFRIs, denoted as $m_I = \frac{\partial w}{\partial q}$. To do this, we first differentiate (2) with respect to q:

$$\frac{\partial \overline{EU}}{\partial q} = \frac{\partial[(1-q)U(\overline{w})]}{\partial q} + \frac{\partial[qI(\overline{w})]}{\partial q}.$$

---

[4] This relative value is also estimated in other works as Persson et al. (2001).

[5] The estimation of the value of a statistical life should also be subjected to a proper method that is different from that used in this study for the estimation of value of preventing a NFRI. For a presentation of this method see Carthy et al (1999).

[6] This variation in wealth can be considered as the Willingness to Pay for reducing the risk of road injury which is estimated by CV and MSG as explained below in section five.

Since the individual expected utility remains constant then $\frac{\partial \overline{EU}}{\partial q} = 0$, and the above expression would be like

$$-U(w) + (1-q)U'(w)\frac{\partial w}{\partial q} + I(w) + qI'(w)\frac{\partial w}{\partial q} = 0;$$

$$\left[(1-q)U'(w) + qI'(w)\right]\frac{\partial w}{\partial q} = U(w) - I(w);$$

$$\frac{\partial w}{\partial q} = \frac{U(w)-I(w)}{(1-q)U'(w)+qI'(w)}.$$

Now this derivative is evaluated at the initial moment at which the individual faces a change in risk of non-fatal injury. That is, we set $q = \overline{q}$ and $w = \overline{w}$, and the MRS is:

$$m_I = \left.\frac{\partial w}{\partial q}\right|_{q=\overline{q}} = \frac{U(\overline{w})-I(\overline{w})}{(1-\overline{q})U'(\overline{w})+\overline{q}I'(\overline{w})}. \tag{3}$$

Analogously if we assume that individuals have a probability $\overline{p}$ of death $\{D, \overline{w}\}$ otherwise having their current state $\{U, \overline{w}\}$, then this lottery can be expressed as $A' = [\overline{p}, \{D, \overline{w}\}; 1 - \overline{p}, \{U, \overline{w}\}]$ and its expected utility is $\overline{EU} = (1-\overline{p})U(\overline{w}) + \overline{p}D(\overline{w})$ where $U(\overline{w})$ and $D(\overline{w})$ are the utility of initial wealth conditional on normal health and death, respectively. It is possible to calculate MRS of wealth for risk of death, m_D, at the baseline moment with risk of death $\overline{p}$. To do this, we first differentiate with respect to (w.r.t.) p, then rearrange, set $p = \overline{p}$ and $w = \overline{w}$:

$$m_D = \left.\frac{\partial w}{\partial p}\right|_{p=\overline{p}} = \frac{U(\overline{w})-D(\overline{w})}{(1-\overline{p})U'(\overline{w})+\overline{p}D'(\overline{w})}. \tag{4}$$

Finally, if we divide (3) by (4) we obtain the ratio of the MRS of wealth for risk of non-fatal accident to the MRS of wealth for risk of death. This is what is known as the relative value of preventing the risk of NFRI. If we assume that $\overline{p} = \overline{q}$ is a very low probability we can assume the two denominators in (3) and (4) to be approximately equal, with a small error, and therefore giving place to the next expression:[7]

$$\frac{m_I}{m_D} = \frac{U(\overline{w})-I(\overline{w})}{U(\overline{w})-D(\overline{w})} \times \frac{(1-\overline{p})U'(\overline{w})+\overline{p}D'(\overline{w})}{(1-\overline{q})U'(\overline{w})+\overline{q}I'(\overline{w})} \cong \frac{U(\overline{w})-I(\overline{w})}{U(\overline{w})-D(\overline{w})}. \tag{5}$$

In section five we explain the two methods for the elicitation of the relative value in expression (5).

[7] For this approximation to not be reasonable the difference between the marginal utility of wealth conditional on both the injury and death, this is $I'(\overline{w}) - D'(\overline{w})$, would have to be extremely and untenable high. Viscusi and Evans (1990) found marginal utility of income conditional on workplace injuries to lie between 77% and 92% of $U'(\overline{w})$. Moreover they found lower marginal utility of income for higher severity. If we consider death as the severest health state we can say that $D'(\overline{w}) < I'(\overline{w})$ and given the reliable assumption that $D'(\overline{w}) \geq 0$ the difference could be at most $I'(\overline{w}) - D'(\overline{w}) = 0.92U'(\overline{w})$. Also, we are interested in the relative value at approximately the average risk of fatal accident for the Spanish population ($\overline{p} = \overline{q} = 0.00015$). Then these numbers would give as a result an upward bias of the relative value estimated by (5) of only 0.014%.

4. Relative Value under Cumulative and Third Generation Prospect Theory

In this section we analyse the theoretical relative value of preventing a NFRI under Tversky and Kahneman (1992) Cumulative Prospect Theory (CPT). There are three main novel features of this theory that we incorporate to the current analysis. These are already present in the early version of Prospect Theory (Kahneman and Tversky 1979) with slight modifications:

- States are considered as gains or losses depending on the individual *Reference Point* (*RP*). In the analysis of the previous section we can say that if an individual preferences are such that $\{U,\overline{w}\} \succ \{I,\overline{w}\} \succ \{D,\overline{w}\}$ and $RP = \{U,\overline{w}\}$, then the injury and death states are consider as losses. In case that $RP = \{I,\overline{w}\}$, then normal health is considered as a gain and death is a loss.
- *Losses loom larger than gains*. This is usually known as *loss aversion*. It entails that the utility decrease when going from $RP = \{U,\overline{w}\}$ to $\{I,\overline{w}\}$ is higher than the utility increase when going from $RP = \{I,\overline{w}\}$ to $\{U,\overline{w}\}$.
- *Probability weighting*. Decision makers do not use objective probabilities to calculate the mathematical expectation of the utility of the states, they rather use decision weights applied to each gain and loss to compute the value of a lottery. A nonlinear function transforms outcomes probabilities into probability weights. Usually this Probability Weighting Function (PWF) is inversed S-shaped giving place to lower and upper subadditivity. This feature is also present in the earlier version of Prospect Theory (Kahneman and Tversky 1979) but it was modified in CPT assuming the probability transformation given by Rank-Dependent Expected Utility (Quiggin, 1982), in order to avoid stochastic dominance violations and extending the theory to prospects with more than two outcomes.[8] Decisions weights are applied separately to gains and losses so that two different PWFs arise and the recognition of the *RP* turns to be crucial for this process.

We take careful attention to the *RP* given its importance in the valuation of lotteries. We then follow the usual assumption in previous CPT implementations that the *RP* is one of the states in the prospects being evaluated (see for example Bleichrodt et al., 2007). Thus, we proceed to calculate theoretical m_I and m_D considering one of the states included in the prospects as the RP. First, respondents are told to assume a probability $\overline{q}$ of having state $\{I,\overline{w}\}$ otherwise having their current state $\{U,\overline{w}\}$. Then subjects are given the chance to reduce their risk of NFRI to q allowing their wealth to be decreased to w. Thus subjects elicit w so that the indifference relation holds between the two prospects, $A = [\overline{q},\{I,\overline{w}\},;1-\overline{q},\{U,\overline{w}\}] \sim B = [q,\{I,w\};1-q,\{U,w\}]$. We consider in the analysis that the states ordering is $\{U,\overline{w}\} \succcurlyeq \{I,\overline{w}\} \succcurlyeq \{D,\overline{w}\}$. Then, for illustrative purposes, the CPT value of prospect A is given by the next expressions depending on the RP:

$$\overline{V}[\overline{q},\{I,\overline{w}\};1-\overline{q},\{U,\overline{w}\}] = \begin{cases} \widetilde{RP} - \lambda W^-(\overline{q})[\widetilde{RP} - I(\overline{w})] - \lambda[1 - W^-(\overline{q})][\widetilde{RP} - U(\overline{w})] & if\ RP \succcurlyeq \{U,\overline{w}\} \succcurlyeq \{I,\overline{w}\} \\ \widetilde{RP} - \lambda W^-(\overline{q})[\widetilde{RP} - I(\overline{w})] + W^+(1-\overline{q})[U(\overline{w}) - \widetilde{RP}] & if\ \{U,\overline{w}\} \succcurlyeq RP \succcurlyeq \{I,\overline{w}\} \\ \widetilde{RP} + [1 - W^+(1-\overline{q})][\widetilde{RP} - I(\overline{w})] + [W^+(1-\overline{q})][U(\overline{w}) - \widetilde{RP}] & if\ \{U,\overline{w}\} \succcurlyeq \{I,\overline{w}\} \succcurlyeq RP \end{cases} \quad . \qquad (6)$$

[8] From now on we shall refer indistinctly to prospect or lottery meaning the same concept.

The specification of the lottery value follows closely that of Bleichrodt et al. (2007), where $\widetilde{RP}$ is the utility of the *RP*, λ is the loss aversion parameter, and W^- and W^+ are the PWFs for losses and gains, respectively.[9] Because we consider the *RP* to be varying we do not follow the standard convention that $\widetilde{RP} = 0$. For the same reason we consider the loss aversion parameter to not be included in the utility function ($U(\overline{w})$ and $I(\overline{w})$) but rather to be a multiplier of utility looses. In CPT practical applications not only it is usual to assume the *RP* to be one of the states in the prospects being evaluated but, moreover, there is empirical support for assuming that the *RP* is one of the states included in the prospect that remains constant to be compared with another lottery whose outcomes are varied by the respondent in order to find the indifference between the two prospects (Stalmeier and Bezembinder 1999, Morrison 2000, Bleichrodt et al. 2001). So in this case the *RP* could be either $\{U, \overline{w}\}$ or $\{I, \overline{w}\}$ in prospect A. The next equation shows the equality of the value of prospect A, on the LHS, and prospect B, on the RHS, when RP=$\{U, \overline{w}\}$:[10]

$$U(\overline{w}) - \lambda W^-(\overline{q})[U(\overline{w}) - I(\overline{w})] = \\ U(\overline{w}) - \lambda W^-(q)[U(\overline{w}) - I(w)] - \lambda[1 - W^-(q)][U(\overline{w}) - U(w)]. \quad (7)$$

Now we can compute MRS of wealth for risk of NFRI by differentiating (7) w.r.t. q. Since we only consider wealth in prospect B to be variable and denoting $W^-{}'(q)$ as the derivative of the PWF for losses, the next expressions hold:

$$0 = \frac{\partial\big[W^-(q)[U(\overline{w}) - I(w)]\big]}{\partial q} + \frac{\partial\big[[1 - W^-(q)][U(\overline{w}) - U(w)]\big]}{\partial q};$$

$$0 = W^-{}'(q)[U(\overline{w}) - I(w)] - W^-(q)\frac{\partial w}{\partial q}I'(w) - W^-{}'(q)[U(\overline{w}) - U(w)] \\ - [1 - W^-(q)]\frac{\partial w}{\partial q}U'(w);$$

$$\frac{\partial w}{\partial q} = \frac{W^-{}'(q)[U(w) - I(w)]}{W^-(q)I'(w) + [1 - W^-(q)]U'(w)}.$$

At the initial situation the MRS of wealth for risk of NFRI is:

$$m_I = \frac{\partial w}{\partial q}\bigg|_{q=\overline{q}} = \frac{W^-{}'(\overline{q})[U(\overline{w}) - I(\overline{w})]}{W^-(\overline{q})I'(\overline{w}) + [1 - W^-(\overline{q})]U'(\overline{w})}. \quad (8)$$

Similarly we are able to compute m_D. Respondents may be required to state their Willingness To Pay for reducing a risk of a fatality from $\overline{p}$ to p so that the next indifferent is obtained, $A' = [\overline{p}, \{D, \overline{w}\}; 1 - \overline{p}, \{U, \overline{w}\}] \sim B' = [p, \{D, w\}; 1 - p, \{U, w\}]$, and assuming RP=$\{U, \overline{w}\}$ then the next equation is true:[11]

$$U(\overline{w}) - \lambda W^-(\overline{p})[U(\overline{w}) - D(\overline{w})] = \\ U(\overline{w}) - \lambda W^-(p)[U(\overline{w}) - D(w)] - \lambda[1 - W^-(p)][U(\overline{w}) - U(w)]\{U, \overline{w}\}. \quad (9)$$

---

[9] The PWF is sometimes denoted by lower-case w. However we reserve the lower-case w for wealth and consider a capital W for the PWF.

[10] In Appendix 1 it is shown that the theoretical relative value obtained in this section does not vary when considering *RP*=$\{I, \overline{w}\}$.

[11] Notice that we follow the empirical findings in the literature that a state seems to be a *RP* only if it appears in the fixed prospect being evaluated (Stalmeier and Bezembinder 1999, Morrison 2000, Bleichrodt et al. 2001). Thus the *RP* is either $\{U, \overline{w}\}$ or $\{D, \overline{w}\}$ in prospect A'.

Differentiating (9) w.r.t. p and evaluating at the initial moment that the respondent face a change in risk of fatality, the MRS of wealth for risk of death is given by the next expression:

$$m_D = \frac{\partial w}{\partial p}\bigg|_{p=\overline{p}} = \frac{W^{-\prime}(\overline{p})[U(\overline{w})-D(\overline{w})]}{W^{-}(\overline{p})D'(\overline{w})+[1-W^{-}(\overline{p})]U'(\overline{w})}. \qquad (10)$$

Now we divide (8) by (10) and obtain the relative value of preventing the risk of NFRI. Again we can say that to a good approximation the two denominators in (8) and (10) are equal (since $\overline{p} = \overline{q}$ is very low) therefore giving place to the next expression:[12]

$$\frac{m_I}{m_D} = \frac{U(\overline{w})-I(\overline{w})}{U(\overline{w})-D(\overline{w})} \times \frac{W^{-}(\overline{p})D'(\overline{w})+[1-W^{-}(\overline{p})]U'(\overline{w})}{W^{-}(\overline{q})I'(\overline{w})+[1-W^{-}(\overline{q})]U'(\overline{w})} \cong \frac{U(\overline{w})-I(\overline{w})}{U(\overline{w})-D(\overline{w})}. \qquad (11)$$

Therefore under CPT the ratio $\frac{m_I}{m_D}$ is approximately the same expression as under EUT (see equation 5 above).

In order to allow for the RP to be a lottery, rather than a certain state, we make use of Third Generation Prospect Theory presented by Schmidt et al. (2008). In PT[3] a lottery X is described as an outcome X_i (e.g. a state as defined previously, $S = \{hp, w\}$) for each *state of the world* (SOW) i that happens with some probability p_i. Now let us define RP as the *reference point lottery* which gives outcome RP_i in SOW i. Then, an outcome given by the lottery X is considered as a loss (gain) if $RP_i - X_i > 0$ ($X_i - RP_i > 0$). We include this property of PT[3] in our framework by generalizing expression (6) to the case for which the reference point is different in each SOW i (this is, the *RP* is a lottery) and there exist more than two SOWs. The value of the lottery X with reference point RP is:

$$V[X, RP] = \widetilde{RP} + \sum_{i=1}^{n} \pi_i^{+}(UT_i - \widetilde{RP}_i) - \lambda \sum_{i=-m}^{-1} \pi_i^{-}(\widetilde{RP}_i - UT_i), \qquad (12)$$

$\widetilde{RP}$ is the value of the *reference point lottery*: $V[RP, RP] = \widetilde{RP}$, that again is not necessarily equal to zero. $\widetilde{RP}_i$ and UT_i are the utilities obtained in the SOW i by the *RP* and X lottery respectively. Where SOWs are ordered in such a way that $UT_i - \widetilde{RP}_i > UT_j - \widetilde{RP}_j$ iff $i > j$. There are $i = 1, \ldots, n$ SOWs with gains indexed with positive numbers and $i = -1, \ldots, -m$ SOWs with losses indexed with negative numbers.[13] The loss aversion parameter is λ. Eventually, π_i^{+} and π_i^{-} are the decisions weights which depend on the probability of SOWs transformed by the PWF for gains and losses as in Tversky and Kahneman (1992):

$$\pi_n^{+} = W^{+}(p_n), \pi_{-m}^{-} = W^{-}(p_{-m})$$
$$\pi_i^{+} = W^{+}(p_i + \ldots + p_n) - W^{+}(p_{i+1} + \ldots + p_n) \; if \; 1 \leq i \leq n-1$$
$$\pi_i^{-} = W^{-}(p_{-m} + \ldots + p_i) - W^{-}(p_{-m} + \ldots + p_{i-1}) \; if \; 1-m \leq i \leq -1.$$

In next we show that the theoretical relative value of preventing a NFRI does not change when considering specification (12). Suppose again that a person is given the opportunity to pay an amount of money in order to have a lower risk of injury, then we will consider the initial situation to be the reference point lottery because it can be considered as the endowment. Thus the reference point will be lottery $A = [\overline{q}, \{I, \overline{w}\}; 1-\overline{q}, \{U, \overline{w}\}]$ in contrast to the new situation

[12] Taking into account the same considerations as in footnote 7 and using the PWF estimated by Tversky and Kahneman (1992) the relative value estimated by (11) would result in an upward bias of only 0.21%.

[13] Notice that states of the world with neutral value ($UT_i - \widetilde{RP}_i = 0$) are not considered in (12) for the obvious reason that it will make no difference in the final value of prospect X.

$B = [q, \{I, w\}; 1 - q, \{U, w\}]$. Naturally, three different SOWs can be considered. The first one is SOW_1 in which lottery B will give the pair of outcomes $\{U, w\}$ and lottery A will give $\{I, \overline{w}\}$. SOW_1 will occur with probability $(\overline{q} - q)$ which is the actual gain that the individual experiences from reducing the risk of injury. Another state of the world SOW_{-1} will occur with probability q in which lottery B will give $\{I, w\}$ and A will give $\{I, \overline{w}\}$. In SOW_{-1} a loss has happened as a consequence of the fact that individual has reduced his wealth but still some degree of risk of injury exists. Finally, state of the world SOW_{-2} occurs with probability $1 - \overline{q}$ where lottery B will give $\{U, w\}$ and A will give $\{U, \overline{w}\}$. There is again a loss since the wealth of the individual is decreased and none of the sources of risk has eventually happened. We assume the utility loss in SOW_{-1} to be lower than in SOW_{-2} given the empirical evidence that $U'(\overline{w}) > I'(\overline{w})$ (see footnote 7) and take that into account when computing decision weights for each state of the world.

The next equation shows the equality of the value of lottery B ($V[B, A]$) according to (12), on the LHS, and the value of the initial lottery A ($V[A, A]$), on the RHS:

$$V[A, A] + W^+(\overline{q} - q) \times [U(w) - I(\overline{w})] - \lambda W^-(1 - \overline{q}) \times [U(\overline{w}) - U(w)] - \lambda[W^-(1 - \overline{q} + q) - W^-(1 - \overline{q})] \times [I(\overline{w}) - I(w)] = V[A, A]. \quad (13)$$

In order to compute the underlying MRS of wealth for risk of injury we derivate (13) w.r.t. q given place to the next expression:

$$W^+(\overline{q} - q)U'(w)\frac{\partial w}{\partial q} - W^{+\prime}(\overline{q} - q)[U(w) - I(\overline{w})] + \lambda W^-(1 - \overline{q})U'(w)\frac{\partial w}{\partial q} - \lambda\left[W^{-\prime}(1 - \overline{q} + q)[I(\overline{w}) - I(w)] - [W^-(1 - \overline{q} + q) - W^-(1 - \overline{q})]I'(w)\frac{\partial w}{\partial q}\right] = 0. \quad (14)$$

Now rearranging, considering the initial situation in which respondents face the trade-off between wealth and risk of injury (this is $\overline{q} = q$ and $\overline{w} = w$), and given the value of the PWF at zero and one ($W^+(0) = 0, W^-(1) = 1$) the MRS is:

$$m_I = \left.\frac{\partial w}{\partial q}\right|_{q=\overline{q}} = \frac{W^{+\prime}(0)[U(\overline{w}) - I(\overline{w})]}{\lambda\left[[1 - W^-(1 - \overline{q})]I'(\overline{w}) + W^-(1 - \overline{q})U'(\overline{w})\right]}. \quad (15)$$

Similarly we are able to compute m_D. Assuming lottery A' to be the reference then the value of A', right hand side (RHS), and B', left hand side (LHS), should be equal to:

$$V[A', A'] + W^+(\overline{p} - p) \times [U(w) - D(\overline{w})] - \lambda W^-(1 - \overline{p}) \times [U(\overline{w}) - U(w)] - \lambda[W^-(1 - \overline{p} + p) - W^-(1 - \overline{p})] \times [D(\overline{w}) - D(w)] = V[A', A']. \quad (16)$$

Differentiating (16) w.r.t p and evaluating at the initial moment that the respondent faces a change in risk of fatality, the MRS of wealth for risk of death is given by:

$$m_D = \left.\frac{\partial w}{\partial p}\right|_{p=\overline{p}} = \frac{W^{+\prime}(0)[U(\overline{w}) - D(\overline{w})]}{\lambda\left[[1 - W^-(1 - \overline{p})]D'(\overline{w}) + W^-(1 - \overline{p})U'(\overline{w})\right]}. \quad (17)$$

Dividing (16) by (17) we obtain the relative value of preventing the risk of NFRI with the next approximation:[14]

[14] Taking into account the same considerations as in footnote 7 and using the PWF estimated by Tversky and Kahneman (1992) the relative value in (18) would result in an upward bias of only 0.30%.

$$\frac{m_i}{m_D} = \frac{U(\overline{w}) - I(\overline{w})}{U(\overline{w}) - D(\overline{w})} \times \frac{[1 - W^-(1-\overline{p})]D'(\overline{w}) + W^-(1-\overline{p})U'(\overline{w})}{[1 - W^-(1-\overline{q})]I'(\overline{w}) + W^-(1-\overline{q})U'(\overline{w})} \cong \frac{U(\overline{w}) - I(\overline{w})}{U(\overline{w}) - D(\overline{w})}. \quad (18)$$

Again when considering a lottery as a reference the theoretical ratio $\frac{m_I}{m_D}$ should be approximately equal to the same expression than under EUT. The fact that expressions (5), (11) and (18) are the same indicates that theoretical relative values has to be interpreted in the same way independently of the assumptions made about the behaviour under risk considered (EUT, CPT or PT[3]). Specifically relative values should be determined by the ratio of the utility decrease in case of injury to the utility decrease with a fatality.

5. The relative value estimation procedures

In this section we show the link between MSG and CV responses and theoretical relative values. The relationship between individual responses and their relative values is the key point for elicitation of preferences.

5.1. Modified Standard Gamble

With the modified standard gamble method health states valuation is done by individuals by making choices between two hypothetical risk situations. One situation, denoted as *Treatment A*, is such that the individual suffers a particular non-fatal injury, which is the one being valued, with the probability $(1-\theta)$ and otherwise (s)he dies, with $\theta > 0$ probability. Another situation, denoted as *Treatment B*, is such that the individual continues with his/her normal health with probability $(1-\pi)$ and otherwise he dies, with $\pi > 0$ probability. The objective is to find a π^* such that the individual is indifferent between this two situations given a level of θ. In the study concerned here the parameter θ is fixed for every choice so the respondent makes repeating choices depending on different levels of π that are suggested in a way such that the conclusion of the valuation of the non-fatal injury implies to obtain the indifference level.

In our survey θ is equal to 0.001 (1 in 1000). An example of a formulated MSG question in the survey is:

> *Suppose that you had a traffic accident and that, in case of not receiving medical care, you could die. There exists two treatments that, in principle, could be applied to your case: Treatment A and Treatment B. Suppose that with treatment A 999 of 1000 people have state V, while 1 in 1000 treated people dies. With treatment B the chances of dying are 400 in 1000 and the chances of returning to their normal health before the accident are 600 in 1000.*

Notice that $\pi = 0.4$ in this example and two possible responses arise. On the one hand, an individual can choose treatment A (treatment B) giving place to another question with $\pi < 0.4$ ($\pi > 0.4$) in order to get closer to the estimation of risk of death in treatment B that makes him indifferent between both situations. On the other hand, the individual can report that both treatments are equally preferred implying that the indifferent risk of death level is 400 in 1000 ($\pi^* = 0.4$).[15]

---

[15] In the next, *Treatment B* will be referred to as the alternative lottery that gives a probability of dying π and a probability of resulting in normal health state equal to $1-\pi$.

Before the valuation questions are presented a training question is included so that the respondents adequately understand and get used to the procedure. In addition to numerical probabilities, visual aids are presented consisting of a panel with represented human figures. The proportion of darkened figures represented the probability. In example of Figure 4, we can see that the respondent has to choose between *"Tratamiento A"* (Treatment A), which entails a probability of 1 and 999 in 1000 of death and injury L respectively, and *"Tratamiento B"* (Treatment B), which gives a probability of one half of death and normal health.

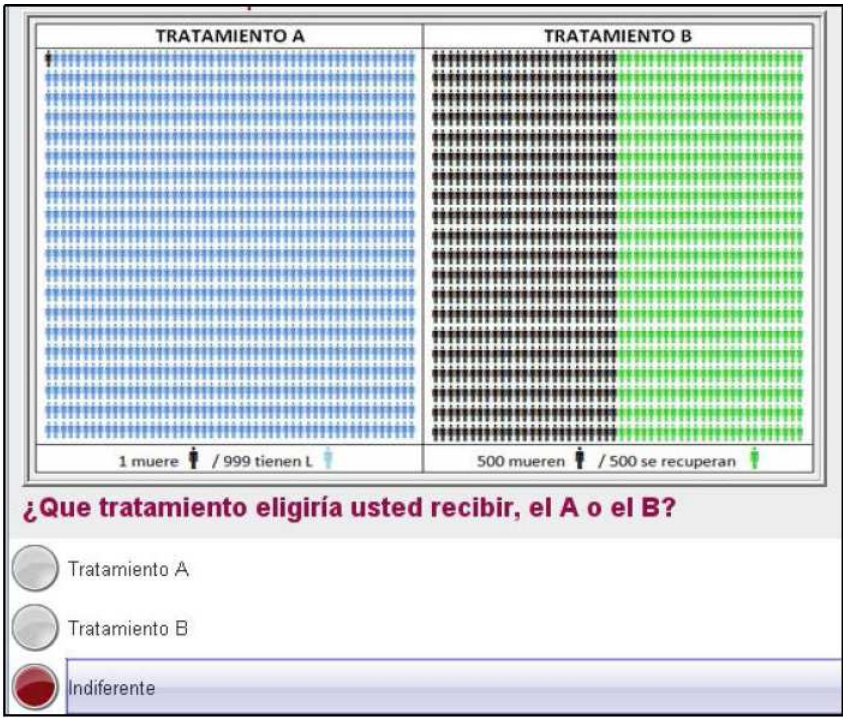


Figure 4. Example of a MSG question on the laptop screen

Given the structure of the MSG questions we can derive standard gamble relative values indirectly from individual responses under EUT, CPT and PT[3] frameworks as shown in the next. We first show the EUT case. At the indifference level, π^*, the expected utility of treatment A, LHS of the equation, equals expected utility of treatment B in the RHS, and the next expression holds:

$$(1-\theta)I(\overline{w}) + \theta D(\overline{w}) = (1-\pi^*)U(\overline{w}) + \pi^* D(\overline{w}). \tag{19}$$

If we add $(\pi-\theta)U(\overline{w})$ in both side of equation (19), then we add $-(1-\theta)I(\overline{w})$ and $-\pi D(\overline{w})$, this expression follows,

$$(\pi^*-\theta)U(\overline{w}) + (1-\theta)I(\overline{w}) + \theta D(\overline{w}) = (1-\theta)U(\overline{w}) + \pi^* D(\overline{w});$$

$$(\pi^*-\theta)U(\overline{w}) + \theta D(\overline{w}) = (1-\theta)[U(\overline{w}) - I(\overline{w})] + \pi D(\overline{w});$$

$$(\pi^* - \theta)[U(\overline{w}) - D(\overline{w})] = (1 - \theta)[U(\overline{w}) - I(\overline{w})];$$

And finally giving equation (20),

$$\frac{\pi^* - \theta}{1 - \theta} = \frac{U(\overline{w}) - I(\overline{w})}{U(\overline{w}) - D(\overline{w})}. \quad (20)$$

Now, by (5) and (20) we have

$$\frac{m_i}{m_D} = \frac{\pi^* - \theta}{1 - \theta}. \quad (21)$$

Eventually (21) connects directly MSG responses, π^*, to the relative value of preventing a risk of non-fatal injury.

However CPT differs from the above result given the alternative interpretation of MSG responses. We compute the MSG relative values under CPT considering varying reference points, $RP = \{U, \overline{w}\}$, $RP = \{I, \overline{w}\}$ and $RP = \{D, \overline{w}\}$. If we consider $\{I, \overline{w}\}$ as *RP* (see Appendix 2 for the analysis with $RP = \{U, \overline{w}\}$ and $RP = \{D, \overline{w}\}$) then the value of *Treatment A*, on the LHS of equation (22), should equal to the value of *Treatment B*, on the RHS, such that the next holds

$$I(\overline{w}) - \lambda W^-(\theta)[I(\overline{w}) - D(\overline{w})] = I(\overline{w}) - \lambda W^-(\pi^*)[I(\overline{w}) - D(\overline{w})] + W^+(1 - \pi^*)[U(\overline{w}) - I(\overline{w})]. \quad (22)$$

Rearranging and scaling, $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$,

$$I(\overline{w}) = \frac{W^+(1-\pi^*)}{\lambda[W^-(\pi^*) - W^-(\theta)] + W^+(1-\pi^*)}. \quad (23)$$

Now if we combine equations (11), scaling $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$, and (23) we obtain the relative value of preventing a NFRI by MSG method under a CPT framework when RP= $I(\overline{w})$,

$$\frac{m_I}{m_D} = \frac{\lambda[W^-(\pi^*) - W^-(\theta)]}{\lambda[W^-(\pi^*) - W^-(\theta)] + W^+(1-\pi^*)}. \quad (24)$$

In the next we analyse MSG responses assuming that the reference is a lottery, as proposed by PT[3], specifically one of the two treatments. When interpreting MSG answers in terms of PT[3] we have to make an assumption about the relationship between the sources of risk in the lotteries between which respondents make a choice. In the case of the MRS analysed in sections 3 and 4 lotteries A and B are correlated because one is just a slightly modification of the other (with lower risk of injury/death and lower wealth). However, in the case of *Treatments A and B* it does not seem plausible to assume that the risks in both lotteries are correlated but rather to assume that both lotteries are independent since they are presented to the respondents as different treatments. When the treatments are independent we can consider four SOWs and the probabilities of each of those can be computed by multiplying the outcome probabilities in both lotteries. In one state of the world treatment A would result in death with utility $D(\overline{w})$ and treatment B would result in normal health with utility $U(\overline{w})$ with probability $\theta(1 - \pi)$. Another state of the world would give $I(\overline{w})$ and $U(\overline{w})$ for treatments A and B respectively, which occur with a chance of $(1 - \theta)(1 - \pi)$. A third state of the world would be when both treatments result in death with utility $D(\overline{w})$ with a probability of $\theta\pi$. Eventually, it could happen that treatment A would give $I(\overline{w})$ and treatment B $D(\overline{w})$ with probability $(1 - \theta)\,\pi$. Now if the reference is treatment A (see Appendix 3 when the RP is treatment B), the MSG answer π^* should be such that the value of both treatments are equal, this is $V[Treat\ B, Treat\ A] = V[Treat\ A, Treat\ A]$:

$$V[Treat\ A, Treat\ A] + W^+(\theta(1-\pi^*))[U(\overline{w}) - D(\overline{w})] + \left[W^+(1-\pi^*) - W^+\left(\theta(1-\pi^*)\right)\right][U(\overline{w}) - I(\overline{w})] - \lambda W^-((1-\theta)\pi^*)[I(\overline{w}) - D(\overline{w})] = V[Treat\ A, Treat\ A]. \quad (25)$$

Rearranging and scaling, $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$,

$$I(\overline{w}) = \frac{W^+(1-\pi^*)}{\lambda W^-((1-\theta)\pi^*) - W^+(\theta(1-\pi^*)) + W^+(1-\pi^*)}. \quad (26)$$

Now if we combine equations (18), scaling $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$, and (26) we obtain the relative value of preventing a NFRI by MSG method under PT[3] with treatment A as reference,

$$\frac{m_I}{m_D} = \frac{\lambda W^-((1-\theta)\pi^*) - W^+(\theta(1-\pi^*))}{\lambda W^-((1-\theta)\pi^*) - W^+(\theta(1-\pi^*)) + W^+(1-\pi^*)}. \quad (27)$$

In Table 3 we can see the computation of the relative value of NFRI depending on the theory and reference point we consider. If we compare expressions in Table 3 we can notice that the sign of the bias between CPT or PT[3], on the one hand, and EUT, on the other hand, is not straightforward. We can try to figure out the bias when $RP = I(\overline{w})$ for example. To do so we have to interpret differences between expression (21) and (24). The factors that makes EUT ratios and CPT ($RP = I(\overline{w})$) ratios to differ are the loss aversion parameter, the specification of W^- and W^+, and individual responses. We arrive to the conclusion that:

- **The effect of the probability weighting, W^- and W^+, is to increase (decrease) CPT ratio with respect to EUT relative value when MSG responses π^* are low (high).** When π^* is low the numerator in (24) is higher than in (21) because given lower subadditivity we have that $W^-(\pi^*) - W^-(\theta) > \pi^* - \theta$. When π^* is high the numerator in (24) could be lower than in (21) given that under upper subadditivity $W^-(\pi^*) - W^-(\theta) < \pi^* - \theta$, however this may be counterbalanced by a high loss aversion parameter.
- **The loss aversion parameter makes CPT ratio to be higher than EUT relative value.** It is easy to see that a higher λ makes the numerator in (24) increase in a higher proportion than it makes the denominator increases. However, loss aversion has no effect in EUT ratios. This is consistent with the fact that when mixed prospects are evaluated then subjects are more risk averse when loss aversion is assumed, as argued in Bleichrodt et al. (2007). Since we are considering RP= $I(\overline{w})$ Treatment B is a mixed prospect, because it contains a gain, $U(\overline{w})$, and a loss, $D(\overline{w})$, and it is the risky prospect with respect to Treatment A that contains only a loss, $D(\overline{w})$, and a state that is neither a loss nor a gain, $I(\overline{w})$. In this context a subject is more reluctant to risk his life in Treatment B, he is not likely to have a high π^*. And if an individual is willing to risk it is so because the utility of the injury in Treatment A has to be very low for him and as a result it happens that the relative value of preventing that NFRI increases (see in equation 11 that relative value increases as $I(\overline{w})$ decreases).

For illustrative purposes we represent the bias between EUT and (C)PT[(3)] in Figure 5 when assuming the PWFs estimated by Tversky and Kahneman (1992),

$$W^{+}(p) = \frac{p^{\gamma}}{[p^{\gamma}+(1-p)^{\gamma}]^{1/\gamma}}; \qquad W^{-}(p) = \frac{p^{\delta}}{[p^{\delta}+(1-p)^{\delta}]^{1/\delta}}. \tag{28}$$

With median $\gamma = 0.61$ and $\delta = 0.69$. They also estimate a median $\lambda = 2.25$.

Table 3. MSG relative value under EUT, CPT and PT[3] with varying RP

		$\boldsymbol{\frac{m_I}{m_D}}$
EUT:		$\frac{\pi^* - \theta}{1-\theta}$
CPT:	$RP = U(\overline{w})$	$\frac{W^-(\pi^*) - W^-(\theta)}{1 - W^-(\theta)}$
	$RP = I(\overline{w})$	$\frac{\lambda[W^-(\pi^*) - W^-(\theta)]}{\lambda[W^-(\pi^*) - W^-(\theta)] + W^+(1-\pi^*)}$
	$RP = D(\overline{w})$	$\frac{W^+(1-\theta) - W^+(1-\pi^*)}{W^+(1-\theta)}$
PT[3]:	$RP = Treat\ A$	$\frac{\lambda W^-((1-\theta)\pi^*) - W^+(\theta(1-\pi^*))}{\lambda W^-((1-\theta)\pi^*) - W^+(\theta(1-\pi^*)) + W^+(1-\pi^*)}$
	$RP = Treat\ B$	$\frac{W^+((1-\theta)\pi^*) - \lambda W^-(\theta(1-\pi^*))}{W^+((1-\theta)\pi^*) + \lambda[W^-(1-\pi^*) - W^-(\theta(1-\pi^*))]}$

Note. Respondents were told to assume $\theta = 0.001$ and π^* is the individual response to MSG. *RP* stands for *Reference Point*.

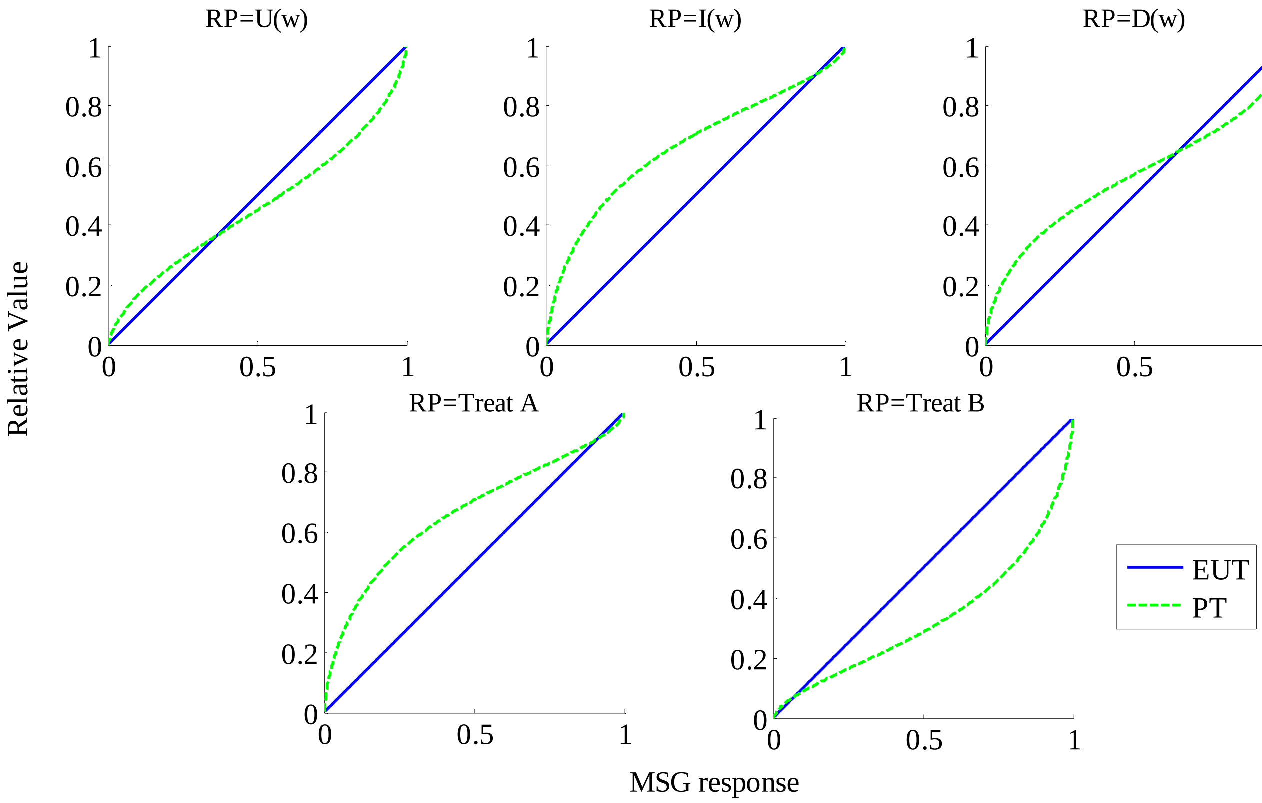


Figure 5. Bias between EUT and PT relative values under different RPs

5.2. Contingent valuation

This method involves directly asking respondents about the amount of money they are willing to pay for reducing the risk of traffic accident. We ask respondents for the quantity to pay for avoid risk of non-fatal accident, on the one hand, and to avoid the probability of death, on the other hand, so we can compute the relative value of preventing a NFRI by dividing both responses since the amount of risk to be reduced is the same in both type of questions.

An example of a formulated CV question in the survey is:

> *Suppose your risk of injury such as W as a result of a traffic accident is 15 in 100,000 and that there exists a safety device that will reduce your risk of health status such as W in a traffic accident in 5 / 100,000, from 15 in 100,000 to 10 in 100,000.*

To represent the baseline risk and the scope of the risk reduction visual aids analogous to that in Figure 4 are used. However, since the amounts of the risks to be represented are small it is difficult their visualization on a laptop screen so that the visual aids were shown in paper.

The CV responses are not open but interactive payments cards are used instead. Each card contains one of the amounts of money shown in Table 4. The way of dispensing the payments cards is as follows. First, all the amounts of money are placed on the screen at random. Then, also randomly, the respondent assesses the quantities one by one indicating in each case whether he/she would "pay for sure" that amount, "not pay for sure" or "not know”. The respondent has to click on the corresponding payment card and take it to the appropriate box according to their response. In Figure 6 it is shown an example in which those amounts that would be paid for sure are supposed to be in the right box, those that for sure would not be paid in the left box, and the rest of the amounts in the box at the bottom. Finally, the exact number that correspond to the maximum WTP is asked by an open question whose responses has to be between the maximum amount in the right hand box and the lowest amount in the left hand box. The computer application is programmed so that respondents could not give inconsistent answers. In the event that this occurs, the respondent is invited to review and correct their answers. Moreover, subjects are said to consider their budget constraints when responding.

Table 4. Amounts of money (€) in the CV payments cards

10	30	50	100	150	300	600
1,000	3,000	6,000	10,000	30,000	100,000	300,000

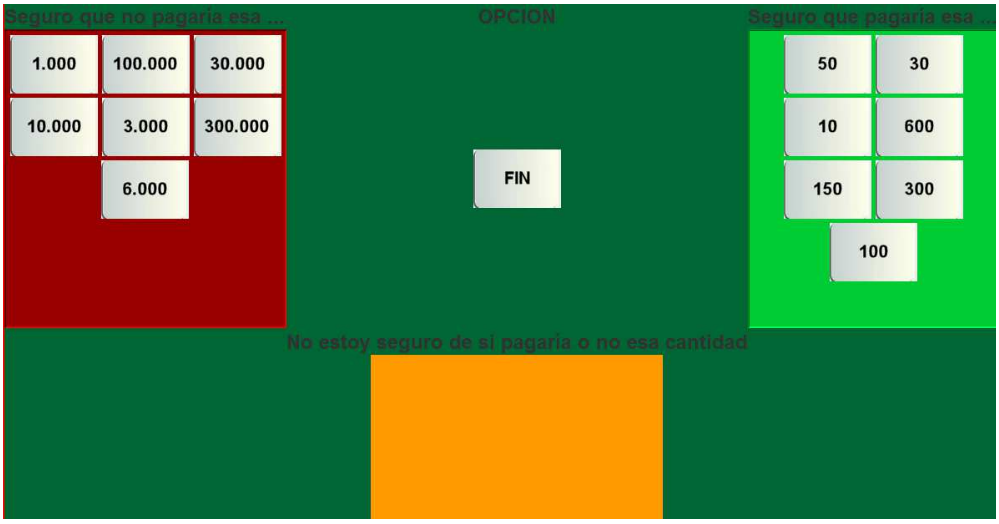


Figure 6. Example of a CV question on the laptop screen

Respondents are told that the security device is for single use. That way we are able to individualize the value of preventing a non-fatal accident and avoid responses taking into account other's people safety. For example, a head of family would think that if (s)he travels often with other household members they also benefit from this safety device and thus be willing to pay higher amounts of money. This feature is crucial when aggregating willingness to pay responses in order to not overstate the value of risk reduction, assuming that improvement in other's people safety is always appreciated. Respondents are also told that this device works in all modes of transport in order to avoid different responses according to their transport habits and make them believe that they would benefit from increased security anyway.

Another aspect that deserves comment is the fact that respondents are told explicitly that the safety device has one year of duration. This latter characteristic of the CV questions is important because it implies that assessment of traffic safety programs has to be made according to how many non-fatal accidents can be avoided in one year and how much society is willing to pay for that.

All CV questions, for the valuation of prevention of NFRIs or a fatality, are made assuming that respondent pay for a risk reduction of 5 in 100,000 accidents (from 15 in 100,000 to 10 in 100,000) so the relative values are not affected by the embedding effect which can appear due to insensitivity of willingness to pay responses to the size of safety improvement.[16]

The MRS of wealth for risk of non-fatal accident is the amount of money a person is willing to give up for an infinitesimal risk reduction. Given that the safety improvement assumed in CV question is sufficiently small we can compute the MRS as the ratio of the amount of money a respondent is willing to pay for the safety improvement, wtp_I, to the risk reduction considered:

[16] The embedding effect appears when a person is willing to pay the same amount of money for two goods that are different in magnitude, for an example see Kahneman and Knestch (1992). Jones-Lee et al. (1995) found considerable insensitivity of WTP responses to changes in risk reduction.

$$m_I = \frac{\partial w}{\partial q}\bigg|_{(p,q)=(\overline{p},\overline{q})} \cong \frac{wtp_I\ (euros)}{risk\ reduction}. \qquad (29)$$

In the same manner the computation of MRS of wealth for risk of death is as follows:

$$m_D = \frac{\partial w}{\partial p}\bigg|_{(p,q)=(\overline{p},\overline{q})} \cong \frac{wtp_D\ (euros)}{risk\ reduction}. \qquad (30)$$

Finally if we divide (29) by (30), and given that risk reduction is the same in both cases, we obtain the relative value of preventing a NFRI estimated by Contingent Valuation as the next ratio:

$$\frac{m_I}{m_D} = \frac{wtp_I\ (euros)}{wtp_D\ (euros)}. \qquad (31)$$

6. Results

6.1. Responses to MSG and CV questions

In Table 5 and Table 6 mean and median responses to MSG and CV questions are shown, respectively. For the MSG responses, it can be observed that as injury severity increases so does the average and median risk taken in treatment B in order to avoid treatment A which includes the injury as an outcome. If we compare mean with median figures it is obvious that responses do not follow a symmetrical distribution. The mean is always bigger than the median for the six less severe injuries (F, W, X, V, S and R). On the contrary, the mean is below the median for the two severest non-fatal injuries (N and L).

In order to avoid extremely conservative responses, when a respondent were willing to take a risk of death lower than 5 in 1000 ($\pi^* < 5$) interviewers were given instructions to suggest him that he should choose a response such that $\pi^* > 1$ for the risk that makes Treatment A and B indifferent.[17] This pattern is of particular concern. For example, in J-L study a classical form of Standard Gamble were carried out to elicit relative values and a huge percentage of respondents did not take any risk to avoid a sure injury (see Table 3 in J-L). In order to compare conservative responses in our study with those in Jones-Lee et al. (1995), we show the percentage of people whose risk taken is lower than or equal to 2 in 1000 in our survey (row five of Table 5). The percentage of people that give conservative responses when valuing W is 53.3% while the corresponding figure with non risk taken responses in J-L is 81.5%. Also for the rest of the NFRIs valued in both, the conservative responses are lower in the present study. For X, S and R the percentages are respectively: 47.5% vs 75.2%; 19.2% vs 41.2%; and 14.5% vs 25.0%.

[17] The reason for this suggestion is that in case of $\pi^* = 1$ *Treatment B* should always be chosen as it is a situation in which the risk of death is the same as in *Treatment A*, since $\theta = 1$, but with the possibility of recover normal health rather than suffer an injury. This argument could not be correct if a respondent prefers a particular injury to his/her normal health. However this kind of preference occurs according to the ranking task for a very low percentage of individuals that goes from 3.6% in the valuation of F, to only 0.9% in the valuation of V.

Table 5. Mean and median responses to MSG questions (π^*)

	F	W	X	V	S	R	N	L
Mean	27.3	31.0	36.6	172.7	211.1	322.0	649.4	815.7
Median	2	2	3	15	25	110	851	960
Observations	1011	1011	1005	1005	1008	1008	1008	1008
$1 < \pi^* \leq 2$	62.1%	53.3%	47.5%	20.9%	19.2%	14.5%	6.0%	3.1%

Note. $1 < \pi^* \leq 2$ refers to the % of total who accept 2 in 1000 or less of risk of death in Treatment B.

For the CV responses the average willingness to pay for the reduction of risk of NRFIs goes from €233.6, for the less severe health profile F, to €15,429, for the severest L, while the median responses for the same injuries goes from €30 to €170, respectively. This pattern reveals two features: first, there is a considerable gap between means and medians figures following a positively skewed distribution; and second, this gap is much higher for the most severe NFRIs. Another characteristic of CV is that some respondents are not willing to pay any money to avoid an injury risk, which is like if they do not value safety at all. If we compare €0 responses shown in Table 6 to the analogous responses in Table 1 of Jones-Lee study, we notice that this pattern is more prominent in the present survey. For example, 16.2% and 12.6% are zero responses for X and R, while in the reference study these figures were just 5% and 3.7% respectively. Finally, in terms of willingness to pay to avoid a risk of death the same feature is found, a positively skewed distribution and large gap between mean and median.[18]

Table 6. Mean and median responses to CV questions (€)

	F	W	X	V	S	R	N	L	D
Mean	233.6	299.7	766.2	1643.3	751.1	1007.1	3496.7	15429.0	31394.3
Median	30	50	50	100	120	150	150	170	160
Observations	758	505	754	749	501	752	503	503	2016
Response=€0	24.8%	20.6%	16.2%	11.5%	10.6%	12.6%	13.9%	14.5%	11.1%

Note 1. Individuals pay for a risk reduction from 15 to 10 in 100,000.
Note 2. Response=€0 refers to the number of people who are willing to pay no money for a risk reduction.

6.2. Relative Value of preventing a Non-Fatal Road Injury

Given the theoretical frameworks outlined above MSG answers are directly connected to individual relative values as shown in equations (21), (24) and (27). Notice that it is not possible to disentangle neither m_I nor m_D without additional information. Therefore aggregation of population relative value of preventing a NFRI is done by computing representative measures as mean or median individual ratios, shown in Table 7. The main objective of this study is that of comparability between the two preference elicitation methods so Table 7 shows the average and

[18] Even though we obtain large differences between median and mean willingness to pay, the fact that the former is lower than the latter is not surprising because consumption is positively related to income which commonly has a positively skewed distribution (see for example Atkinson, 1995, and Neal and Rosen, 2000).

median individual ratios based on CV answers and on MSG responses under different theories of decision under risk, EUT and CPT with varying reference points.[19]

Mean and medians relative values estimated by MSG are lower than by CV. The gap is even wider for average estimations. In absolute terms the mean and median gaps are respectively increasing and decreasing with severity. In relative terms, by contrast, the gap is sharply declining with respect to severity. This is illustrated by the fact that the average relative value of preventing a risk of F, the mildest NFRI, is 17 times higher by CV, and the value of preventing a risk of the severest one, L, by CV is only 2.7 times MSG estimation under EUT.

Table 7. Estimated mean and median individual ratios ($\frac{m_i}{m_D}$) by MSG and CV

		F	W	X	V	S	R	N	L
MSG ratios									
EUT		0.026	0.030	0.036	0.172	0.210	0.321	0.649	0.815
		0.001	*0.001*	*0.002*	*0.014*	*0.024*	*0.109*	*0.851*	*0.960*
CPT	RP=$\{U, \overline{w}\}$	0.031	0.036	0.045	0.176	0.209	0.309	0.606	0.770
		0.005	*0.005*	*0.009*	*0.044*	*0.064*	*0.173*	*0.716*	*0.868*
	RP=$\{I, \overline{w}\}$	0.053	0.063	0.078	0.271	0.312	0.423	0.714	0.847
		0.012	*0.012*	*0.022*	*0.099*	*0.143*	*0.356*	*0.876*	*0.943*
	RP=$\{D, \overline{w}\}$	0.047	0.055	0.068	0.229	0.265	0.367	0.651	0.797
		0.012	*0.012*	*0.022*	*0.089*	*0.125*	*0.284*	*0.768*	*0.880*
PT[3]	RP=Treat A	0.057	0.067	0.082	0.274	0.315	0.426	0.716	0.848
		0.016	*0.016*	*0.026*	*0.103*	*0.147*	*0.359*	*0.877*	*0.943*
	RP=Treat B	0.019	0.022	0.028	0.122	0.149	0.238	0.528	0.708
		0.001	*0.001*	*0.004*	*0.024*	*0.035*	*0.094*	*0.572*	*0.788*
CV ratios									
		0.45	0.56	0.48	0.68	1.28	1.52	1.69	2.24
		0.20	*0.30*	*0.33*	*0.55*	*0.69*	*1.00*	*1.00*	*1.00*

Note 1. Mean ratios shown in standard font and median ratios in italics.
Note 2. CV ratios have been calculated including all observations in Table 6 minus those that have missing values due to €0 wiliness to pay response for D.

Relative values for PT in Table 7 are computed considering the loss aversion parameter and probability weighting function estimated by Tversky and Kahneman (1992). In the next section we consider the estimation of optimum parameters that are the best fit to our data. Mean PT

[19] In Appendix B in Chilton et al. (2002) it is set out five desirable properties to the relative value obtained from matching data methods and it is proposed a consistent way of aggregation. However, this alternative way is only applicable to CV but not to MSG. For example, one of the properties is that of symmetry which is not met by the mean of individual ratios, since the inverse of the average of the individual ratios is not equal to the average of the inverse of the individual ratios. By contrast, median of the individual ratios, also presented in this study, does meet this property.

relative values are more similar to CV ratios than EUT ones when the assumed reference point in MSG responses is either $\{I, \overline{w}\}$, $\{D, \overline{w}\}$, or Treatment A (except for the relative value of L in case $RP = \{D, \overline{w}\}$). The approximation of PT under $RP = \{I, \overline{w}\}$ or $RP = Treat\ A$ is very similar and better than when considering other reference point because it gives place to higher relative values given a specific MSG response as Figure 5 shows (section five). The key is that when a person takes one of these two RPs Treatment B loses desirability, with respect to the EUT case or PT with other RPs, and a specific MSG response is considered as a decrease of $I(\overline{w})$ and hence an increase of the relative value as shown in (5), (11) and (18). Nonetheless, PT modelling of decision under risk does not solve the huge discrepancies emerged in this study between both preference elicitation methods. As can be seen in Table 7 CV relative values more than double those of PT.

We now consider the possibility that the found gap in the estimation of relative values is caused by the fact that MSG is designed to provide the valuation of injuries better than or equally preferred to death and so it only provides relative values lower or equal to the unity. This is not always the case for our respondents, for instance 40.5% and 59.5% of respondents have ranked N and L, respectively, as worse than death. Furthermore, in fact CV could estimate bigger than unity ratios. To the extent that a specific injury is worse than death there is a downward bias in MSG relative values. In order to account for this effect we have computed relative values only for respondents showing CV ratios lower or equal to one. In Figures 7 and 8 it is shown mean and median, respectively, CV and MSG estimations under EUT and PT with different reference points. It can be seen that disparities lessen for all NFRIs, especially for N and L. However, CV ratios continue being bigger by far for the majority of HSs.

Neither within group comparison leads to different results. In Figure 9 and 10 we focus on respondents who answer to the evaluation of the same injuries with the two methods. This implies taking into account only MSG ratios of groups 5 to 8 (see Table 1). Wilcoxon signed rank tests reject equality of CV and MSG ratios for every injury at 1% level of significance.

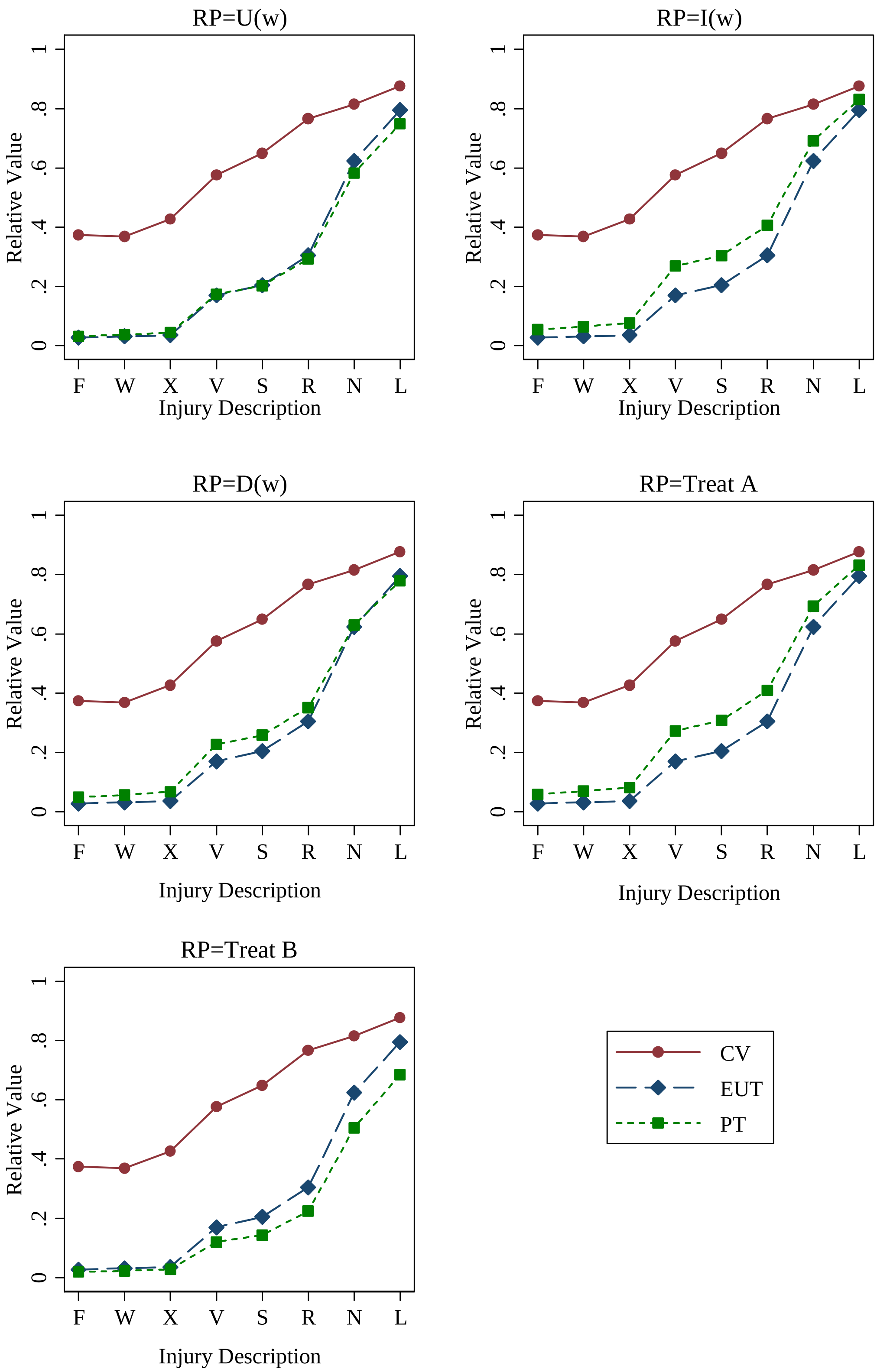


Figure 7. Relative Values when CV ratios are ≤ 1. Mean

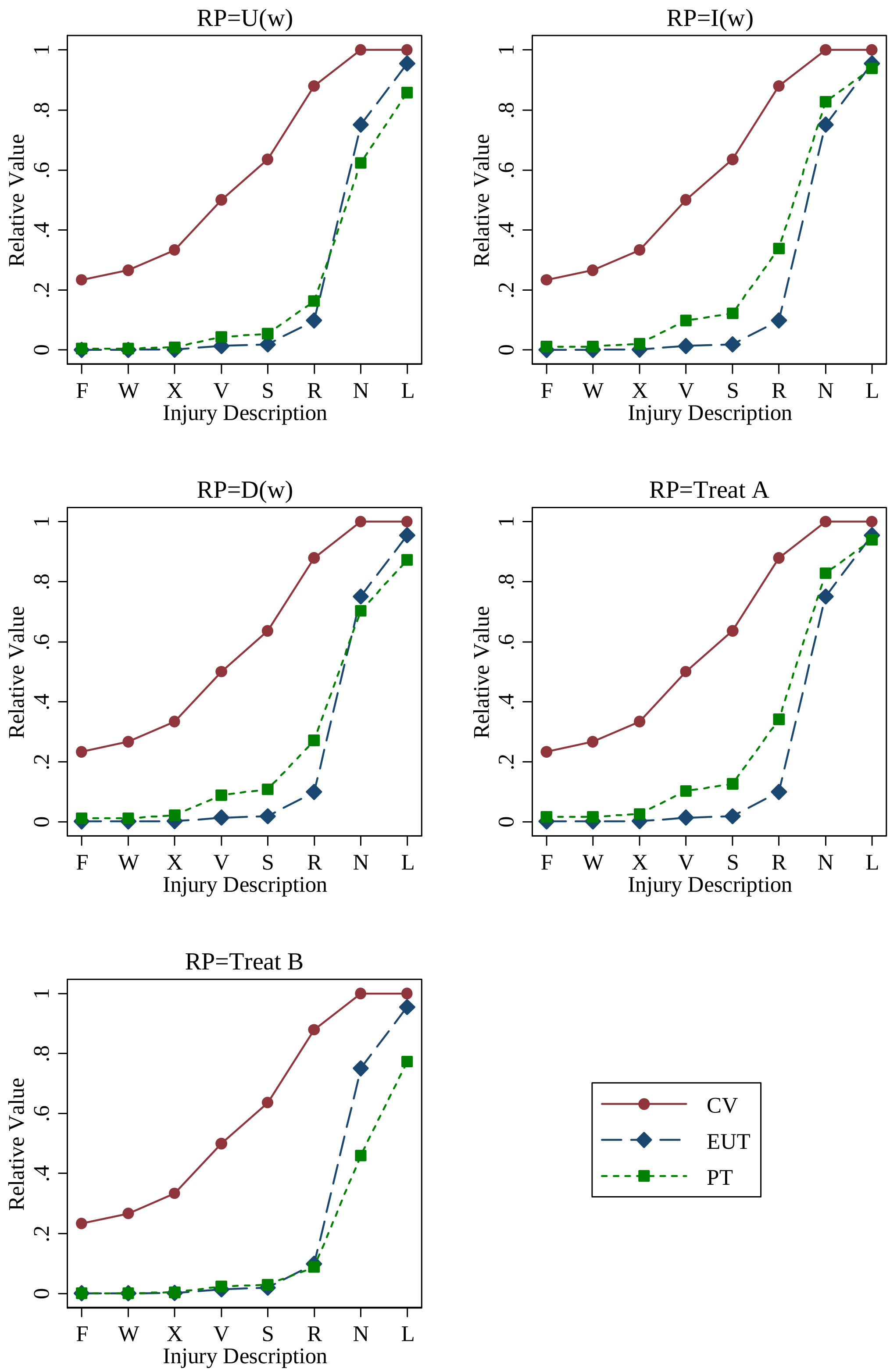


Figure 8. Relative Values when CV ratios are ≤ 1. Median

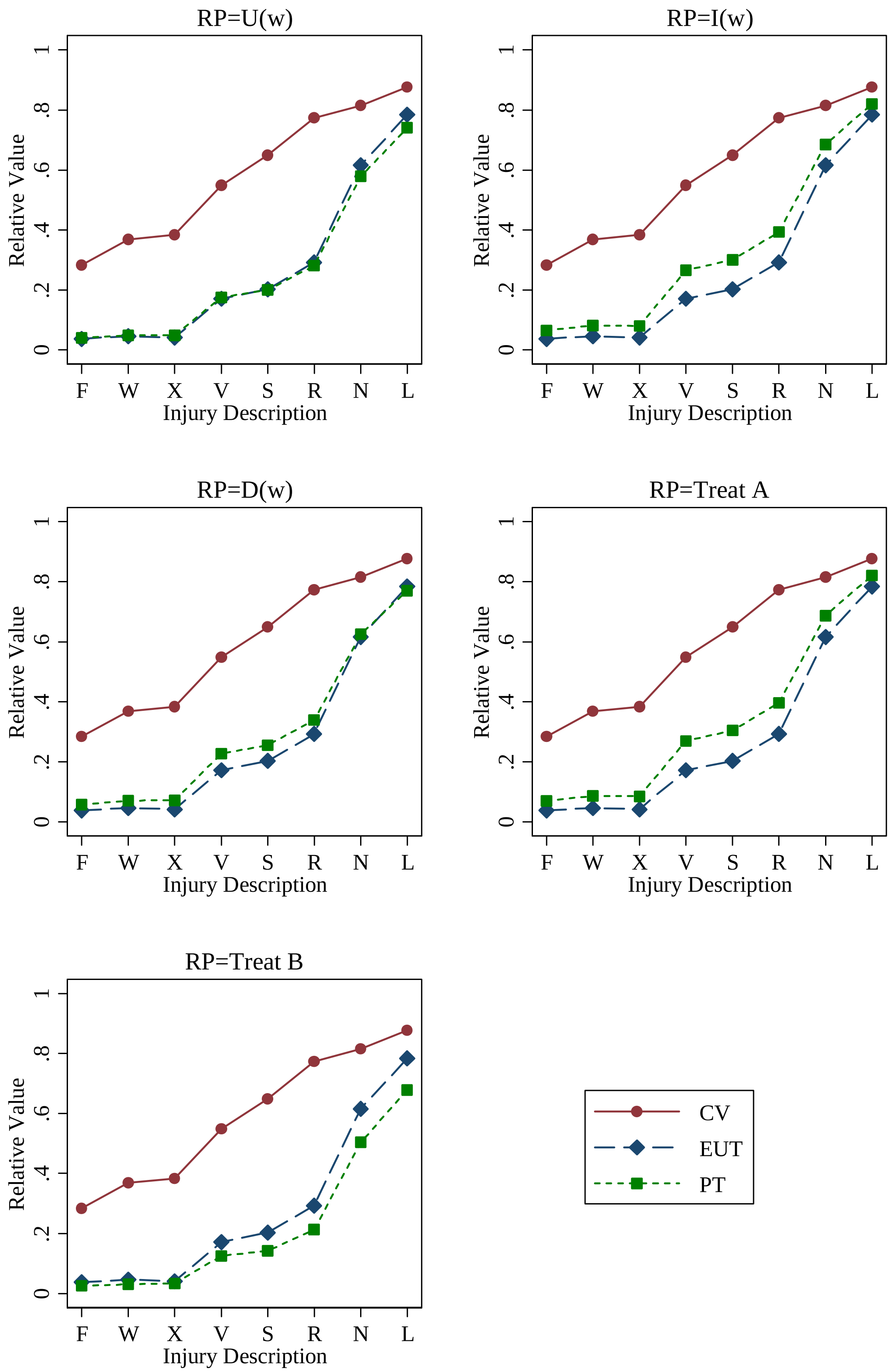


Figure 9. Relative Values when CV ratios are ≤ 1. Within group comparison. Mean

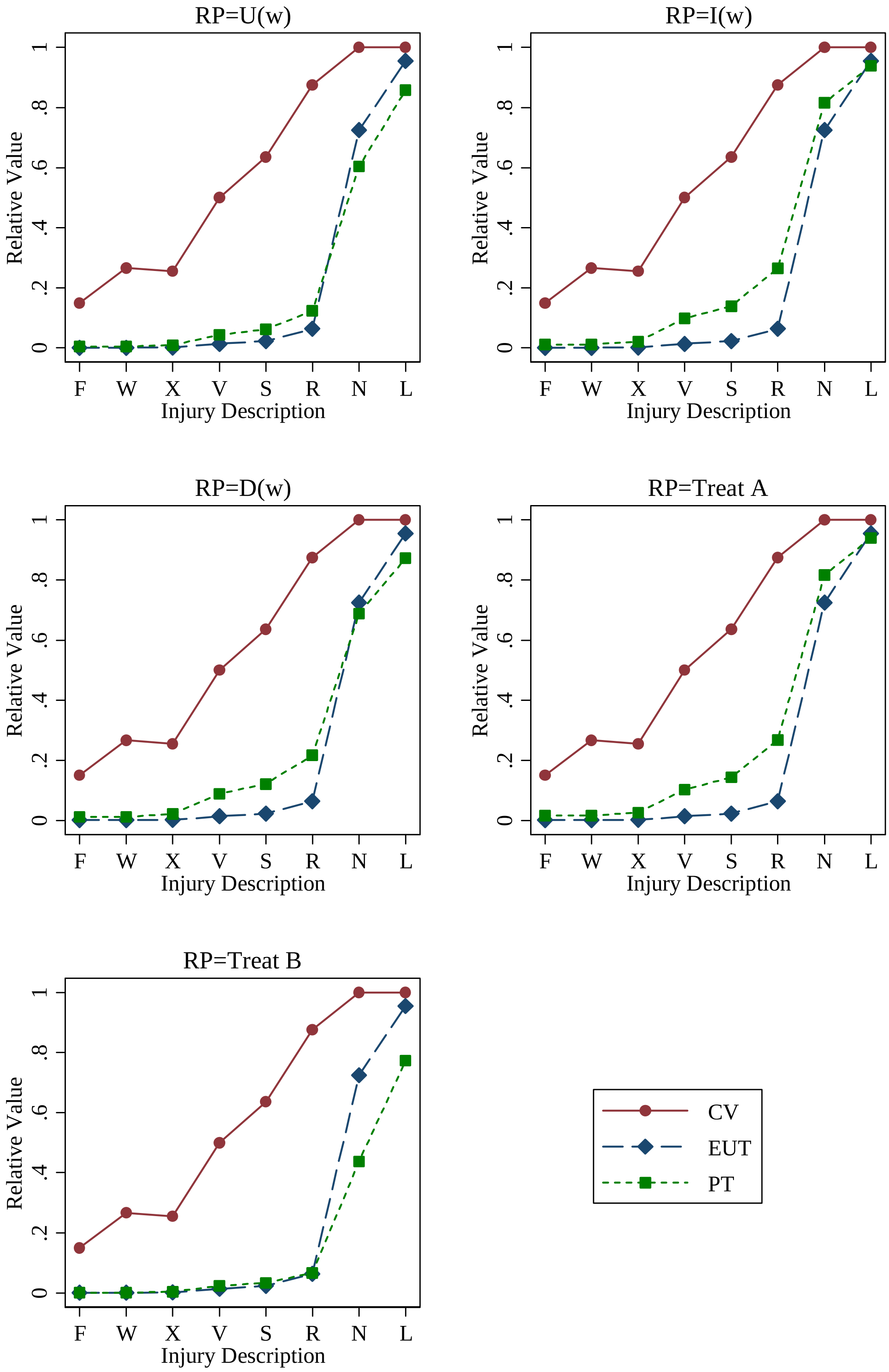


Figure 10. Relative Values when CV ratios are ≤ 1. Within group comparison. Median

6.3. Making Prospect Theory optimum

In this section we analyse whether estimation of optimum parameters for PT could solve the mismatch between CV and MSG. It could be the case that varying the loss aversion parameter λ and the parameters γ and δ of the probability weighting functions in Tversky and Kahneman (1992), we obtain closer relative values. For that we restrict analysis to the reference points with the best approximations previously found: $RP = Treat\ A$ and $RP = \{I, \overline{w}\}$. Parameters are estimated to minimize Mean Absolute Error (MAE) between CV and MSG ratios. This is to minimize $\frac{1}{T}\sum\left(\frac{\boldsymbol{m_I}}{\boldsymbol{m_D}}(CV) - \frac{\boldsymbol{m_I}}{\boldsymbol{m_D}}(MSG)\right)$ where T is the number of individual ratios considered (number of individual multiplied by number of injuries valued). The same parameters are considered for the eight NFRIs and only ratios lower or equal to unity are taken into account to make CV and MSG ratios comparable (see discussion in section 6.2). The estimated parameters in the case of $RP = Treat\ A$ (Third Generation Prospect Theory) are $\lambda = 8.83$, $\gamma = 1.11$ and $\delta = 0.41$. In the case of $RP = \{I, \overline{w}\}$ (Cumulative Prospect Theory) estimations are 10, 0.16 and 0.25, respectively.[20] Minimum MAEs are 0.2901 and 0.2926 for $RP = Treat\ A$ and $RP = \{I, \overline{w}\}$, respectively. Therefore, both PT^3 and CPT MSG ratios implies practically identical approximations to CV ratios.

Mean and Median PT ratios based on optimum parameters are shown in Figure 11 jointly with CV ratios and MSG relative values based on EUT for the within group sample (groups 5 to 8). It can be seen graphically that PT with optimum parameters makes a much better fit to CV than EUT. This is for both CPT (with $RP = \{I, \overline{w}\}$) and PT^3 (with $RP = Treat\ A$). Nonetheless we reject equality of distribution for PT and CV ratios at 1% of error (Wilcoxon signrank tests) for all the injuries except for injury V (p-value=0.318 for $RP = Treat\ A$ and p-value=0.014 for $RP = \{I, \overline{w}\}$). This means that in spite of aggregate (mean and median) PT relative values are quite similar to CV ones, at the individual level when considering the whole distribution they are not the same thing.

The acceptance of PT as a plausible explanation for the discrepancy between MSG and CV ratios relies on the parameters considered for the computation of relative values. Then a relevant question is ¿are optimum parameters reliable? The answer should be made on the grounds of predictability and descriptive capacity in other similar settings. In this sense loss aversion parameter estimated here for CPT (with $RP = \{I, \overline{w}\}$) is about four times higher than those estimated in other implementations of the same version of prospect theory (Tversky and Kahneman, 1992; Bleichrodt et al., 2001; Bleichrodt et al., 2007) implying huge departure from previous empirical evidence. Also, we find higher overweighting (underweighting) for low (high) probabilities. For example, if we consider the probability of death in treatment A, $\theta = 0.001$, the transformed probability is much higher here than in Tversky and Kahneman (1992): $W^+_{\delta=0.25}(\theta) = 0.092 > W^+_{\delta=0.69}(\theta) = 0.008$. This fact is consistent with Rottenstreich and Hsee (2001) that find higher probability weighting for affect-rich outcomes (as the case of health). Eventually, we have to be cautious in the comparability with other studies were lotteries

[20] Estimation was performed using nonlinear optimization with MATLAB 7.11 based on the Nelder-Mead simplex algorithm. In the case of $RP = \{I, \overline{w}\}$ we constrained the loss aversion parameter to be lower than or equal to 10 because unconstrained optimization led to an extremely and unusual huge number, $\lambda = 64.06$. Anyway both, unconstrained and constrained optimum parameters implied similar MAE of 0.2904 and 0.2926, respectively. In addition, we fit our data using the probability weighting functions proposed by Prelec (1998) and Gonzalez and Wu (1999) and the MAEs obtained are very similar. For example, in the case of $RP = Treat\ A$, MAE=0.2893 for both Prelec (1998) and Gonzalez and Wu (1999) functions.

have probabilities higher than 1%, but in contrast our lottery probabilities are very low (θ and π^*). For example, more than 90% of MSG responses are $\pi^* < 0.01$ in the case of injuries F and W.

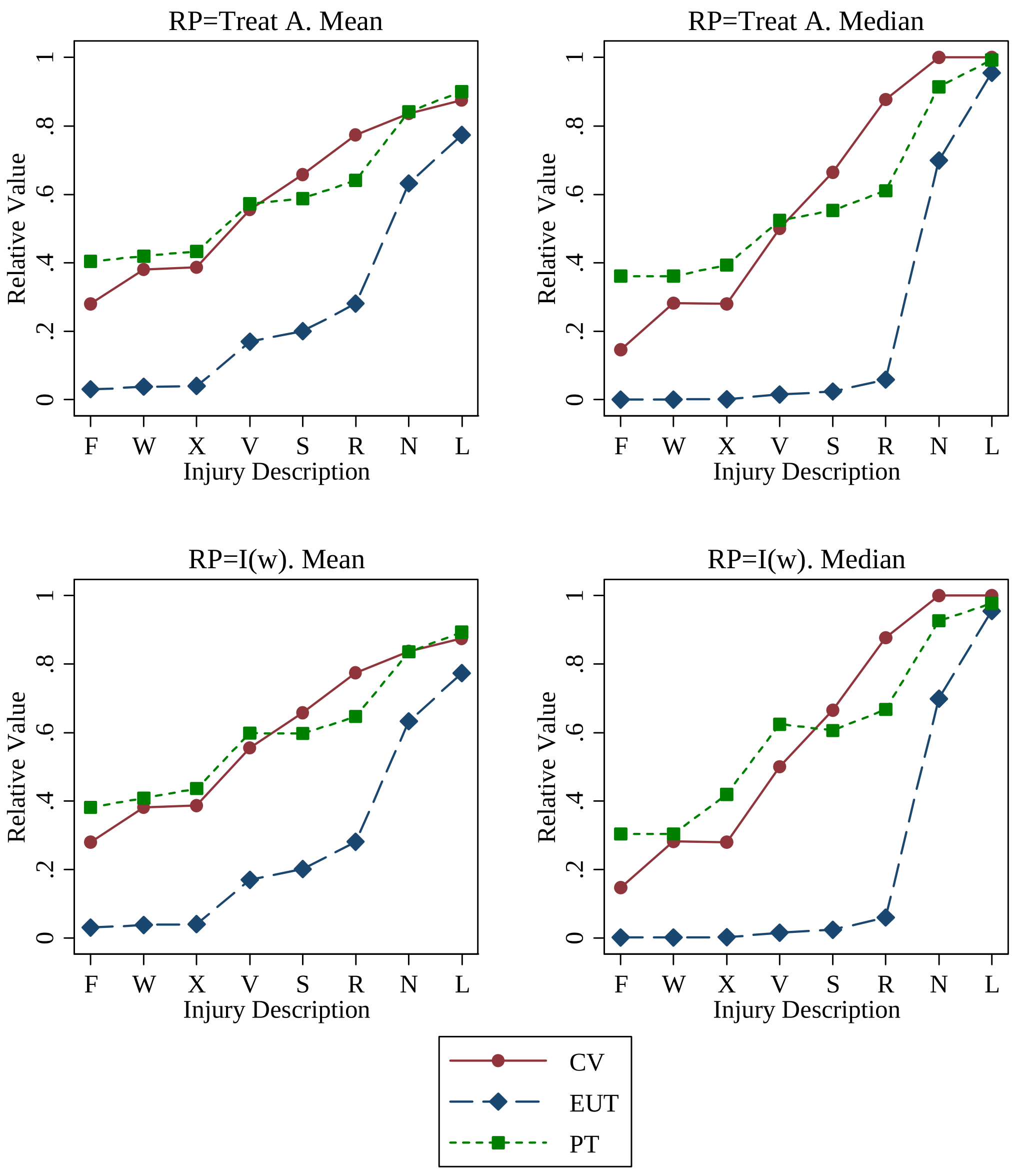


Figure 11. Relative values (CV ratios ≤1) with optimum parameters for PT. Within group comparison

6.4. Considering Visual Analogue Scale scores as utilities

Eventually we consider the possibility that relative values could be estimated by VAS responses. Since respondents assign a score to each NFRI, as well as to death and their normal, health in order to reflect the strength of their preferences between different HSs, this can be interpreted as a way to elicit utilities $U(\overline{w})$, $I(\overline{w})$ and $D(\overline{w})$. Then we can compute relative values through equation (5). It is widely accepted that VAS scores are not utilities of the kind that predict decisions under risk (Parkin and Devlin, 2006), however we use VAS values *as if* they were utilities.

In the two upper graphs of Figure 12 it is displayed within group comparison (groups from 5 to 8) of mean and median ratios for VAS and CV. The same ratios when considering only respondents with ratios lower or equal to unity are shown in the two graphs at the middle. In principle CV ratios bigger than the unity do not have to be a methodological problem for the VAS to replicate that results since ratios bigger than one are allowed in the analogue scale. However observation of mean figures suggests that there is a limitation of the VAS to replicate the high *bigger than one CV ratios*. We believe that it is the lower and upper bounds, 0 and 100, which are behind the explanation because the not permitted lower than zero scores could have been a constraint to reflect the strength of preferences between the worse than death NFRIs and a fatality. Notice that in CV responses there is no upper bound, what allows respondents to answer a much higher willingness to pay for avoiding a risk of injury than for avoiding a risk of death. Finally VAS ratios computed from VAS scores adjusted from Parducci's Range-Frequency effects (R-F, see Parducci, 1965, and Wedell and Parducci, 1988) are displayed in the two graphs at the bottom. The use of this model to adjust VAS values has been done previously in the literature in the context of injuries and health states valuation (Robinson et al., 2001 and Schwartz, 1998). See Appendix 4 for a description of the adjustment procedure.

It can be noticed that VAS methods fits CV estimation much better than MSG under both theories analysed above, EUT and PT. Nonetheless Wilcoxon signed rank tests reject equality of distribution of 6 ratios: at 10% level of significance for injury V; at 2% for injuries W, R and N; and at 1% for injuries X and L. No significant differences are found between CV and VAS ratios for the health profile F and S. When considering only ratios lower or equal to the unity sign rank test cannot reject equality of F, V, N and L ratios. Surprisingly, adjusted VAS that is supposed to be free of context effects is the worst fit of CV although we cannot reject the equality of relative values for injury S.

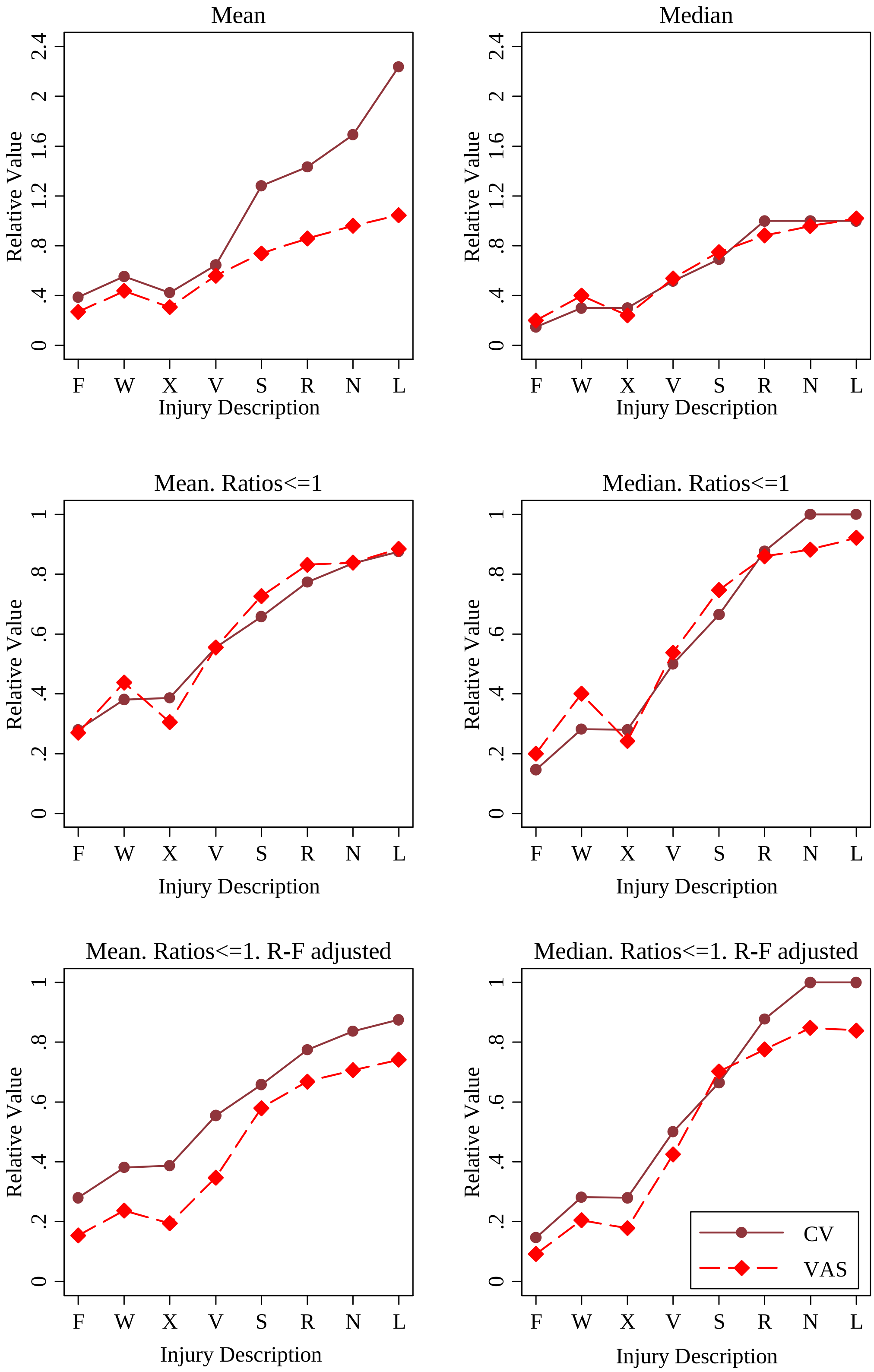


Figure 12. VAS and CV relative values. Within group comparison

7. Consistency analysis of MSG and CV responses

Given the differences found between the two main elicitation procedures included in our survey, we conduct the rest of the analysis to consider criteria that can help us to assess the relative performance of these preference based methods. Specifically, this section is intended to study the consistency of the responses in both procedures.

Since respondents order NFRIs according to their preference in the ranking task, we are able to know whether this order of preference do not change when the same individuals face CV or MSG questions. Therefore, for each pair of health states we know whether the stated preference ordering is consistent with the established order in the ranking. We differentiate two level of consistency: we say a respondent is strictly *consistent* (C) when the strict order over two NFRIs by the stated preferences methods is the same as in the ranking. Likewise, *weakly consistency* (WC) arises when the weak order is the same as in the ranking. Notice that C implies WC but the inverse is not true. Also notice that a non weakly consistent person is making a greater mistake than a non strictly consistent person. This is because the former is declaring preferences that are just the opposite of those indicated in the ranking task, and we will say that this subject is *inconsistent*. For example, if an individual ranking response denotes $F \succ W$, then (s)he is C only if his/her responses in CV or MSG denote the same preference ($F \succ W$), WC if his/her responses denote $F \sim W$ or $F \succ W$, and inconsistent if his/her responses denote $F \prec W$. Eventually, notice also that in the ranking task it is only possible a strictly preference relation whereas in CV and MSG is also possible an indifference relation, so we can distinguish between C and WC.[21]

The consistency of an individual in their responses to MSG (CV) can be expressed by the following binary variable, C_{ij}^{MSG} (C_{ij}^{CV}) which takes the value 1 if individual i is consistent in the ordering of the pair of HSs j and 0 otherwise. We can also express whether an individual is weakly consistent by the next binary variable WC_{ij}^{MSG} (WC_{ij}^{CV}). Therefore, we can perform the following test of proportions which null hypothesis is that percentages of consistent individuals in the valuation of the pair j by MSG and by CV are the same. That is,

$$H_o: \mu_{C_j}^{MSG} = \mu_{C_j}^{CV}$$

$$H_1: \mu_{C_j}^{MSG} \neq \mu_{C_j}^{CV}$$

where $\mu_{C_j}^{MSG}$ and $\mu_{C_j}^{CV}$ are the population means of C_{ij}^{MSG} and C_{ij}^{CV} respectively. So that if the normal distributed z statistic is significant at 1%, 5%, or 10%, we reject the null hypothesis and we conclude that both methods generate a different level of consistency. Similarly, we realize this test for the equality of the percentage of weakly consistent individuals ($H_o: \mu_{WC_j}^{MSG} = \mu_{WC_j}^{CV}$).

[21] We could consider at least three reasons for not being (strictly) consistent or weakly consistent. First, respondent may make mistakes when revealing her/his true preferences in ranking, on the one hand, or in CV or MSG questions, on the other hand. Second, a person could change his preferences, for example because in the course of the interview (s)he changes her/his mind. And finally it could be that the elicitation procedures do not facilitate her/him giving a consistent response. To illustrate the third reason notice that for CV questions, if an individual is not (strictly) consistent it does not mean that he makes a mistake. To see this consider a respondent with no disposable income at all and suppose (s)he has a preference ranking such that $F \succ W$. Imagine that, given his income, her/his willingness to pay for a risk reduction of both injuries is €0, denoting $F \sim W$. This individual is not (strictly) consistent but neither makes a mistake nor changes her/his preferences.

Another issue different from the previous one is testing on possible correlation between consistency in both methods. For that purpose we perform the Chi-square test of independence. The null and alternative hypotheses are:

$$H_o: C_{ij}^{MSG} \text{ and } C_{ij}^{CV} \text{ are independent}$$

$$H_1: C_{ij}^{MSG} \text{ and } C_{ij}^{CV} \text{ are not independent}$$

If the statistic χ^2 is significant at 1%, 5% or 10% of error we can reject the hypothesis of independence at that level of significance. Likewise, we can test the null hypothesis $H_o: WC_{ij}^{MSG}\ and\ WC_{ij}^{CV}\ are\ independent$ to test the independence of weakly consistency in both methods. In addition, we would like to know the strength and sign of the correlation. To do this we calculate the Spearman correlation coefficient between C_{ij}^{MSG} and C_{ij}^{CV}. We also calculate the correlation coefficient between WC_{ij}^{MSG} and WC_{ij}^{CV}.

The above analysis includes a measure of consistency for each health states pairs. However, we are also interested in an aggregated measure of the consistency of each individual. To do this we create an index that consists on adding the number of times respondents are consistent in the ordering of different pairs. Since respondents value four NFRIs plus death the combinatory number of 10 different pairs of health conditions arises and the consistency index for each respondent is computed as: [22]

$$CONS_i^{MSG} = \sum_{j=1}^{10} C_{ij}^{MSG} \text{ and } CONS_i^{CV} = \sum_{j=1}^{10} C_{ij}^{CV}$$

where $CONS_i^{MSG}$ ($CONS_i^{CV}$) is the consistency index that could have any value between 0 and 10. Seemingly we compute the weakly consistency index as:

$$WEAK_i^{MSG} = \sum_{j=1}^{10} WC_{ij}^{MSG} \text{ and } WEAK_i^{CV} = \sum_{j=1}^{10} WC_{ij}^{CV}$$

We first comment on these aggregate consistency indexes. In Table 8 strict consistency and weak consistency indexes are presented for both MSG and CV responses for subgroups 5 to 8.[23] It is seen that the number of consistencies is higher in MSG responses. In average 7.3 out of 10 HSs pairs have been order consistently, i.e. the same order of preference arises in ranking responses and in MSG. The same figure is significantly lower in CV responses (W-test z=9.25), with 5.9 mean consistent pairs. For the weak consistency index results are the other way around, weak consistency is significantly higher in CV responses (W-test z=-14.28). Two comments can be drawn from this analysis. First, the average number of inconsistent responses (non weakly consistent responses) is very low in both methods. Secondly, the comparative evaluation of both methods relies on the concept of consistency we use. If a concept of strict consistency is considered, then MSG is listed as the best performing method. The opposite conclusion arises if we base on a concept of weak consistency.

We compute the Spearman correlation coefficient between $CONS_i^{MSG}$ and $CONS_i^{CV}$, on the one hand, and between $WEAK_i^{MSG}$ and $WEAK_i^{CV}$, on the other hand. Significant positive correlation

[22] In the ranking exercise six health states are valued, four injuries plus normal health and death. However since we do not have CV of normal health (it makes no sense to pay for avoid a risk of being in normal health) only five health states are considered in the consistency analysis.

[23] Notice that computation of CV (weakly) consistency index is only possible for subgroups 5 to 8 since subgroups 1 to 4 value only one NFRI.

of 0.13 and 0.06, respectively, are found. This result has two interpretations: first, a significant positive correlation implies that there are factors that make a person be (weakly) consistent, or not (weakly) consistent, regardless of the method chosen; and second, the fact that the correlation is much lower than unity indicates that there are factors affecting the consistency of each method independently. Later in this section, we do econometric analysis to find some of these factors related to the consistency of responses.

The correlation between (weak) consistency indexes is also illustrated in Figure 13. In panel a) mean $CONS_i^{CV}$ is shown to be increasing with $CONS_i^{MSG}$ and the slope seems to be linear. Panel b) also shows a positive relationship between mean $WEAK_i^{CV}$ and $WEAK_i^{MSG}$ when the latter is lower than 6, however for $WEAK_i^{MSG} > 6$ the slope of the curve is zero, which means no correlation at all.

Table 8. Mean and median consistency and weakly consistency indexes

Indexes	Mean	Median	W-test (z)	Spearman
$CONS_i^{MSG}$	7.3	8	9.25***	0.13***
$CONS_i^{CV}$	5.9	7		
$WEAK_i^{MSG}$	9.0	9	-14.28***	0.06**
$WEAK_i^{CV}$	9.6	10		

Note 1. $CONS^{MSG(CV)}$ and $WEAK^{MSG(CV)}$ are the number of consistent and weakly consistent responses to MSG (CV) questions.
Note 2. The number of observations is 1001 from group 5-8 (see Table 1). Three respondents give a missing response to at least one of the CV questions and are omitted from the analysis.
Note 3. Spearman correlation and equality of distribution Wilcoxon test shown. ***, **, and * denote statistical significance at 1%, 5% and 10% of error respectively.

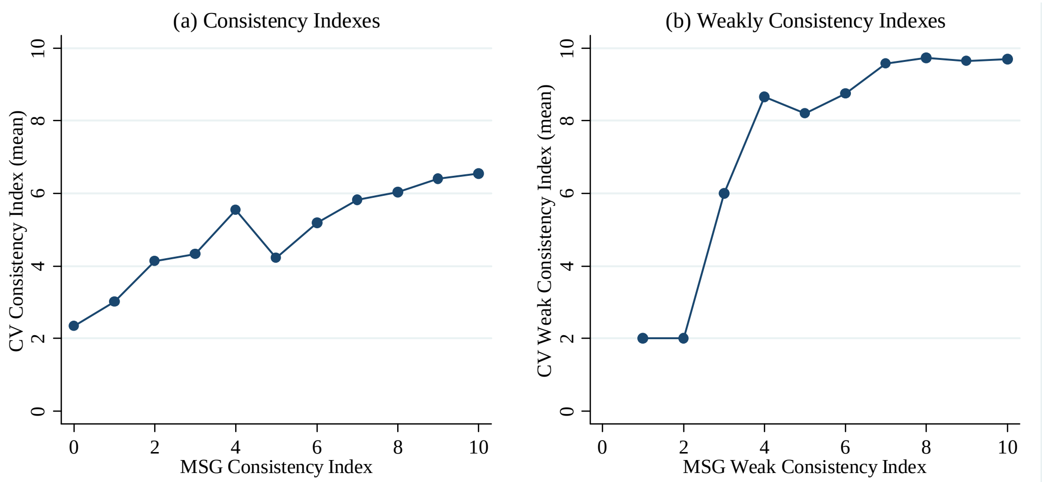


Figure 13. Mean of CV (Weakly) Consistency Index by MSG (Weakly) Consistency Index

In addition, we can study the relationship between the two procedures with contingency tables. To do so we have categorized each index into three subgroups with approximately the same relative frequency in Table 9. Respondents are categorized into three groups according to the number of consistencies: 0 to 6; 7 to 8; and 9 to 10. The percentage of total respondents between 9 and 10 consistencies in CV is 31.2%, but the same proportion is only 23.1, for the subgroup that has between 0 and 6 consistencies in MSG, and as high as 36.7, for the subgroup with 9 or 10 consistencies in MSG. Therefore again we find that the number of consistencies in both methods have a positive relationship. Table 10 shows the proportion of respondents in three categories of $WEAK_i^{CV}$ by $WEAK_i^{MSG}$. We see that a less strong relationship than in Table 9 arises. The proportion of respondent with 10 weak consistencies in CV is somewhat bigger for the subgroup with 10 weak consistencies in MSG, 85.1%, than for the subgroup with 0 to 8 weak consistencies in MSG, 81.4%.

Table 9. Consistency indexes contingency table. Three categories. Row percentages

		CV Consistency Index (0 to 10)			
		0 - 6	**7 - 8**	**9 - 10**	**Total**
MSG Consistency Index (0 to 10)	**0 – 6**	52.8	24.2	23.1	100
	7 – 8	44.0	23.5	32.6	100
	9 – 10	38.1	25.3	36.7	100
	Total	44.7	24.2	31.2	100

Note 1. Cut points are chosen to categorize indexes into three groups with similar frequency.
Note 2. The number of observations is 1001 from group 5-8 (see Table 1). Three respondents give a missing response to at least one of the CV questions and are omitted from the analysis.

Table 10. Weak consistency indexes contingency table. Three categories. Row percentages

		CV W. Cons. Index (0 to 10)			
		0 - 8	**9**	**10**	**Total**
MSG W. Cons. Index (0 to 10)	**0 – 8**	12.5	6.1	81.4	100
	9	9.8	5.7	84.4	100
	10	8.8	6.1	85.1	100
	Total	10.1	6.0	83.9	100

Note 1. Cut points are chosen to categorize indexes into three groups. Due to the positive skewness of the Contingent Valuation Weakly Consistency Index distribution, the three groups are far from having the same frequency.
Note 2. The number of observations is 1001 from group 5-8 (see Table 1). Three respondents give a missing response to at least one of the CV questions and are omitted from the analysis.

In the analysis done so far in this section indexes have been used to assess the aggregate consistency in the ordering of the 10 combinations of health profile pairs each respondent has valued together. From now on we study consistency separately for each pair of health states. Table 11 shows the percentage of consistent individuals in both methods for each of the 28 different pairs of health states ordered by the respondents.[24] The percentage figures that are significantly higher are in bold. We can say that both methods achieve varying degrees of consistency, since most proportions tests results in the rejection of the null hypothesis. For 17 pairs, out of 28, percentage of consistent individuals is greater in MSG. Only three pairs of health states have been ordered more consistently in CV. For the rest couples of health conditions no significant differences between the two methods are found. These results clearly imply that in general MSG produces more strictly consistent responses.

Table 11. Percentage of consistent responses in CV and MSG responses by NFRIs pairs

	F			W			X			V			S			R			N			L		
	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.
W	**49.3**	21.9	***																					
X																								
V							53.2	**63.5**	***															
S	**70.1**	57.7	***	**63.2**	54.2	**	67.7	70.2		45.6	42.7													
R	70.5	64.4		66.8	64.8		73.0	78.6		54.8	57.3		46.9	50.7										
N	72.7	**86.2**	***	68.8	**86.2**	***	70.0	**89.6**	***	62.8	**81.2**	***												
L	73.5	**89.7**	***	72.6	**89.7**	***	73.2	**93.6**	***	67.2	**90.8**	***							41.0	**52.7**	***			
D	70.8	**97.6**	***	67.7	**98.4**	***	71.7	**98.4**	***	61.8	**98.0**	***	54.5	**95.6**	***	46.3	**88.6**	***	34.4	**59.2**	***	40.6	37.4	

Note 1. Proportion test shown (PT). ***, **, and * mean significant differences found at 1%, 5% and 10% of error respectively.
Note 2. MSG (CV) percentages are in bold if significantly higher than CV (MSG) ones.

In any case, both methods seem to generate more consistent responses as the severity gap between the two health states increases. For example, the percentage of consistent respondents for the pair VX is 53.2% and 63.5% for CV and MSG respectively, while the same figures for the DX pair are 71.7% and 98.4% respectively. The increase of the severity gap is captured by the strength of preference shown in VAS responses. This pattern is illustrated in panel a) and b) of Figure 14 where each pair of health states is represented by the proportion of strictly consistent responses in the vertical axis, and average absolute VAS score differences between the two health conditions, in the horizontal axis. A positive relationship arises between consistency and strength of preferences, for CV and MSG, nonetheless, the dispersion of the latter scatter plot is higher which denotes a weaker correlation.

[24] For each respondent we can study the consistency in relation to 10 different health conditions. However as valued NFRIs are not the same for all the groups we can study up to 28 different pair comparisons.

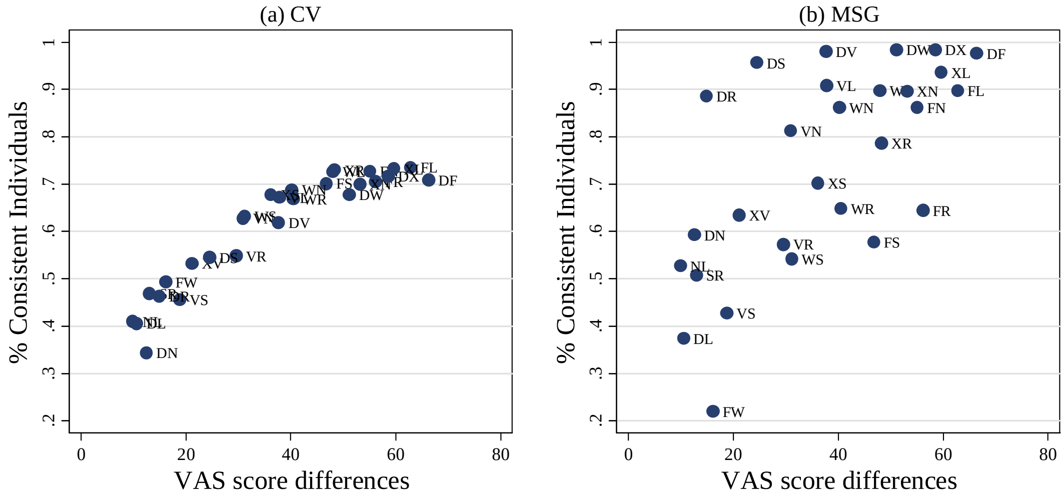


Figure 14. Relation between consistency responses and Absolute differences in VAS score of NFRIs pairs

In Table 12 we show results of weak consistency by each pair of health states. The weak consistency proportions are bigger than 98% and 93%, respectively for CV and MSG, for the vast majority of these pairs. This table also shows that the percentage of weakly consistent individuals is significantly higher for CV in 20 out of 28. If we put this data in evaluation along with Table 11 we conclude that MSG performance has mixed results: on the one hand, this method produces more strictly consistent individuals, on the other hand, more inconsistencies, non-weak consistencies, arise from responses. In contrast, for CV inconsistencies are much lower although it is not the best method to produce strictly consistent responses.

Table 12. Percentage of weakly consistent individuals in CV and MSG by NFRIs pairs

	F			W			X			V			S			R			N			L		
	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.	CV	MSG	P.T.
W	**98**	91.5	***																					
X																								
V							**98.2**	92.2	***															
S	**98.8**	93.3	***	**98.4**	92.1	***	**98.8**	94	***	**96**	79	***												
R	**98.8**	92.5	***	**98.8**	92.9	***	**98.8**	93.5	***	**98**	82.7	***	**97.2**	83.2	***									
N	**98.8**	96.4	*	**99.2**	96.4	**	98.4	97.2		**98**	91.2	***												
L	**99.2**	97.2	*	**100**	97.6	**	98.4	96.8		**98.4**	95.2	**							**97.6**	88.1	***			
D	96.6	**99.6**	***	95.8	**98.6**	***	96.8	**98.6**	*	94.6	**98.6**	***	90.2	**96**	***	89	89.8		**91.5**	70	***	**90.7**	55.9	***

Note 1. Proportion test results shown (PT). ***, **, and * meaning significant differences found at 1%, 5% and 10% of error respectively.
Note 2. MSG (CV) percentages are in bold if significantly bigger than CV (MSG) ones.

As for strictly consistent responses, weak consistency has a positive relationship with the strength of preferences between the two health states. This is the higher the distance in subjective severity the lower the inconsistency. Figure 15 shows a positive relationship between proportion of weakly consistent respondent and average VAS score differences between the two health states. The slope of the scatter plot seems to be larger for MSG.

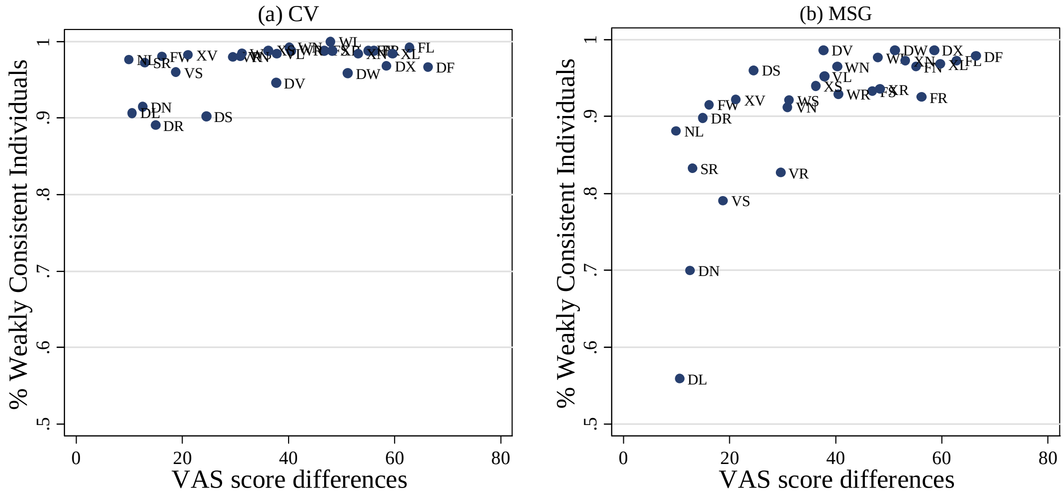


Figure 15. Relation between Weakly Consistent responses and Absolute differences in VAS score of NFRIs pairs

In the next, we analyse correlation of strict and weak consistency between MSG and CV by each pair of HSs. In Tables 13 and 14 it is shown spearman correlation coefficients and independence tests results. For 19 out of 28 pairs the null hypothesis of independence is rejected and a positive relationship is found between C_j^{MSG} *and* C_j^{CV}. Though correlation is not very high it goes from 0.1, for SW, to 0.31, for RV. Significant and negative correlation arises only for pair DL. Also weakly consistent responses in both methods are significantly correlated in 19 out of 28 with even higher spearman coefficients, from 0.1 to 0.49. Again weak consistency seems to be negative correlated for the pair DL.

Results given by independence test of strict consistency is illustrated in contingency tables embedded in Table 16 in Appendix 5, in which correlation between the two categorical variables C_j^{MSG} *and* C_j^{CV} is shown for the 28 pairs of injuries. For example, the highest correlation coefficient is found for the responses to VR, 0.31, and we see that the percentage of strictly consistent responses in CV is higher for consistent respondents in MSG ($C_{j=VR}^{MSG} = 1$), 68.3%, than for non consistent subjects ($C_{j=VR}^{MSG} = 0$), only 36.8%. Contingency tables for weak consistency analysis are embedded in Table 17 (Appendix 5). As an example, we see that the proportion of weak consistent responses in CV for the pair XR is 81.2% when $WC_{j=XR}^{MSG} = 0$, but the same figure is 100% for weakly consistent respondents in MSG ($WC_{j=XR}^{MSG} = 1$).

Table 13. Correlation between consistency in MSG and CV responses

	F	W	X	V	S	R	N	L
W	0.05							
X								
V			0.07					
S	0.08	0.1*	0.17***	0.16**				
R	0.15**	0.13**	0.28***	0.31***	0.14***			
N	0.14**	0.08	0.18***	0.14**				
L	0.21***	0.11*	-0.01	-0.05			0.13***	
D	0.24***	0.15***	0.16***	0.11***	0.14***	0.05	-0.04	-0.17***

Note 1. Spearman correlation coefficients between $C_j^{MSG}\ and\ C_j^{CV}$ shown.

Note 2. Chi2 independence test results shown as ***, **, and * meaning rejection of the null hypothesis at 1%, 5% and 10% of error respectively.

Table 14. Correlation between weakly consistency in MSG and CV responses

	F	W	X	V	S	R	N	L
W	-0.04							
X								
V			0.07					
S	0.26***	0.2***	0.28***	0.1				
R	0.38***	0.4***	0.42***	0.16**	0.02			
N	0.37***	0.22***	0.36***	0.16**				
L	0.26***	------	0.34***	0.27***			0.06	
D	0.1***	0.48***	0.49***	0.32***	0.1**	0.02	-0.01	-0.12***

Note 1. Spearman correlation coefficients between $WC_j^{MSG}\ and\ WC_j^{CV}$ shown.

Note 2. Chi2 independence test results shown as ***, **, and * meaning rejection of the null hypothesis at 1%, 5% and 10% of error respectively.

Note 3. Due to no inconsistent individuals for the pair WL it is not possible the computation of a correlation coefficient.

After comparative analysis of the consistency of the two methods, the remainder of this section is intended to identify factors that may be associated with consistency in the answers. This is possible thanks to collected socio demographic information about respondents. We estimate both a random and fixed individual-specific effect logit model that predicts the probability of a consistent response in CV (MSG) as [25]

$$\Pr\left(C_{ij}^{CV(MSG)} = 1 \middle| \boldsymbol{x_i}, \boldsymbol{\beta}, VASdif_{ij}, \delta, \alpha_i\right) = \frac{\exp\left(\alpha_i + \boldsymbol{x_i'}\boldsymbol{\beta} + \delta VASdif_{ij}\right)}{1 + \exp\left(\alpha_i + \boldsymbol{x_i'}\boldsymbol{\beta} + \delta VASdif_{ij}\right)}. \quad (32)$$

Where $\boldsymbol{x_i}$ is a vector of individual variables that varies between respondents and $\boldsymbol{\beta}$ is the corresponding coefficients vector. $VASdif_{ij}$ refers to absolute VAS score differences, between the two health conditions included in the pair j valued by respondent i, which corresponding

[25] For a detailed explanation of random and fixed effects logit models see Cameron and Trivedi (2005).

coefficient is δ. Notice that VAS score has variation between and within individual variation. Finally, α_i is the individual-specific effect that captures respondent unobserved heterogeneity. The random effect models assumes a normal distribution of this parameter, with $\alpha_i \sim \mathcal{N}[0, \sigma_\alpha^2]$. Nonetheless fixed effect model allows us to estimate δ even controlling by unobserved individual heterogeneity.

Table 18 (Appendix 5) contains the regression analysis of the strict consistency in CV and MSG. It shows 4 different specifications. We first comment on CV results. Specifications 1 to 3 refer to estimates of random effects with different regressors, and specification 4 is the fixed effect estimation of VAS strength of preferences. Specification 1 includes a first analysis with age, education and employment status as explanatory variables. Probability of consistent response in CV is significantly negative related to age and being unemployed and positively related to education. However only age remains significant when controlling in specification 2 for household income, which is very significant. In specification 3 we test the effect of several variables that are collected in the questionnaire. A high smoking frequency is positively related to consistency, but those who smoke very much (>20 cigarettes a week) are not significantly different to non-smoker. Those who practice sports and play gambling games are significantly more consistent. Also respondents who drive a car result to be more consistent. We build a happiness index from 6 questions related to life satisfaction, two of these are: *I am satisfied with my life* and *Conditions of my life are excellent*. Seven possible responses that vary from 1, "*completely disagree*", to 7, "*completely agree*". The average of the six life satisfaction questions is computed and results positively correlated to consistency. Another factor included in the regression is the ability to understand and interpret probabilities discriminating between those who do understand the likelihood of road accident and those who do not. Those people with correct answer to the "test question" (see section 2.3) do not have a better strict consistency. Eventually, the strength of preferences between two health states is included in quadratic form with $VASdif$ and $VASdif$^2 that are highly significant for both random, specification 3, and fixed effect, model 4. The positive coefficient of the former indicates more consistent responses the higher VAS score differences, and the negative coefficient of the latter reduce the positive effect as VAS gap increases. For MSG consistency, variables that are significant once controlling for the complete set of explanatory variables, in equation 3, are: on the one hand, age (only 31-45), working in the public sector, with a negative effect; and on the other hand, household income, less significant than for CV regression, and VASdif, with a positive effect.

Eventually Table 19 shows the results of the same previous analysis for weak consistency in CV and MSG. Again the same 4 specifications as in Table 18 are considered. For CV, explanatory variables that are common in the three equations of random effects do not change, age is significant (only >75 with positive coefficient), and education with a very limited negative effect on weakly consistency (only slightly significant secondary education). Neither age nor education is significant for weak consistency in MSG. Household income is significant for CV but not for MSG in specification 3. For the rest significant variables in explaining CV weak consistency are smoking frequency (only 11-20 cigarettes a week), playing gambling games (high significant and negative effect), and the very high significant quadratic form of VAS gap, the higher this gap the more weak consistency. For MSG understanding the risks arises as a significant factor that increases weakly consistent responses. Also in this method $VASdif$ is the most significant and important variable.

Finding theoretical explanations to the relationships found in Table 18 and 19 is not straightforward and may be adventurous. However there are three variables that we consider interesting in the frame of our analysis: VAS score differences; household income; and understanding of risk. First of all, *VASdif* is systematically positive and significant for strictly and weakly consistency in both CV and MSG. The econometric analysis suggests also a concave relationship given the negative sign of the squared *VASdif*^2. This result challenges stochastic representations of choice based on a constant probability of preference reversals, a constant *error rate* like suggested by Harless and Camerer (1994). On the other hand, it is more in line with Fechner type stochastic models as in Hey and Orme (1994) that predicts that preference reversals are negatively related to the strength of preferences between the alternatives included in the choice set. There are also implications for preference elicitation for similar health states. Special care has to be taken since the error of reporting the true underlying preferences is higher in those cases. The other two variables that deserve comments are income household and understanding of risk. The significant positive effect of the former on consistent responses suggests that insensitivity in CV responses may be caused by budget constraint problems. For the results on the latter variable, people are more weakly consistent (or less inconsistent) in MSG responses when they have previously demonstrated a correct interpretation of risks and probabilities. It seems reasonable to consider to what extent this elicitation methods could be highly sophisticated for some people in order to elaborate appropriate procedures. We further discuss these issues in section 8.4.

8. Discussion

8.1. The mismatch between CV and MSG. The theory

Huge differences arise between the estimated ratio $\frac{\boldsymbol{m_I}}{\boldsymbol{m_D}}$ by both methods, MSG and CV, as in previous surveys. The modified version of the standard gamble does not work to make the gap with the contingent valuation method be reduced although it has reduced the percentage of conservative responses with respect to the standard version used in the *British Study* (compare Table 5 in this study with Table 3 in Jones-Lee et al., 1995). Higher CV relative values of preventing a NFRI are estimated; especially they are relatively higher the less severe the injury is.

The use of Cumulative and Third Generation Prospect Theory as the basis for modelling respondent behaviour has reduced the gap between CV and MSG, so this theory is more consistent with data. Given that CV and MSG can be considered as risk-risk elicitation methods, respondents choose between two risky prospects, this result differs from that encountered by Bleichrodt et al. (2007) where no improvement over expected utility is found for that cases. They explained that by suggesting that there was less distortion in probability weighting in the interval [0.10, 0.20] than previously reported in the literature. Our results support this idea because differences between MSG relative values under EUT and PT do not depend on the probability weighting in that specific interval. As it can be seen in Table 3 for the cases of $RP = I(\overline{w})$ and $RP = Treat\ A$ our results rely mainly on probability transformation of either very low probabilities, like θ or π^* (mostly lower than 0.01), or on very high probability like $(1-\pi^*)$, $(1-\ \theta)$, $(1-\theta)\pi^*$ and $\theta(1-\pi^*)$ (mostly higher than 0.99). Eventually, even though PT gain descriptive power, huge differences still remain for the estimation of relative values between CV and MSG, when considering parameter previously estimated by Tversky and

Kahneman (1992). We have to use optimum parameters for CPT or PT^3 to make MSG and CV ratios similar. Even in that case Wilcoxon signrank tests reject the equality of distribution of MSG and CV ratios for most of the injuries (except for V) which suggests that Prospect Theory cannot explain the whole gap.

Furthermore, Figure 12 and Wilcoxon test show that a better fit of CV relative values is obtained if we use VAS scores as utilities. This best fit is at an aggregate level, mean and median, however it also occurs at an individual level. In top panel of Table 15 we show the proportion of individual responses for which VAS relative values are the best fit of CV estimations.[26] The corresponding percentages for the rest approaches considered in this study are also shown. VAS proportions of best fit are the highest for every NFRIs with the exception of injury X. For that injury the *best fit* is PT^3 with optimum parameters when the reference point is Treatment A ($RP = Treat\ A*$). In order to know how good the fit of the VAS and other approaches is we also compute the Mean Absolute Error (MAE) of VAS estimation as $\frac{1}{T}\sum\left(\frac{\boldsymbol{m_I}}{\boldsymbol{m_D}}(VAS) - \frac{\boldsymbol{m_I}}{\boldsymbol{m_D}}(CV)\right)$, where T is the number of individuals, and compare with the same MEA for the rest approaches. Under this criterion VAS is the closer estimation of CV relative values with the lower MAE for the four severest NFRIs (S, R, N and L). However for three injuries (W, X and V) the lowest MAE is for $RP = Treat\ A*$. This analysis again suggest that VAS score is the best approximation to CV ratios followed closely by $RP = Treat\ A*$. In fact when we compare only these two approaches we obtain that VAS is the best fit for majority of respondents only for injury F, S, R and L. Nonetheless, notice that row VAS scores are a good approximation to CV ratios by themselves, while MSG under PT^3 has to be adjusted to minimize MAE.

Thus VAS is more like CV than MSG under EUT and PT here considered. We think this is evidence for the two methods to be based on a very similar valuation process. For example, CV and VAS valuations have in common that they are as a result of assigning numbers, either an amount of money or a point along a scale respectively, to each health profile. On the contrary, MSG is a choice based method. Eventually, VAS lack of a theoretical connection as that presented here for CV and MSG under both EUT and CPT. As argued in Parkin and Devlin (2006), VAS scores are not utilities despite the fact that VAS has been shown to have interval scale property typical of measurable value functions (see Jones-Lee et al., 1994). The reason why not using VAS scores in Cost Utility Analysis in the health domain is that they do not predict *actual* individual choices under uncertainty and many decision problems are of that kind.

The results presented in Table 15 also indicate that PT outperformed EUT. The MAE is higher for EUT than for PT when considering $RP = I(\overline{w})$, $RP = D(\overline{w})$, and $RP = treat\ A$ (except for injury F). Indeed when the reference point is the injury (in CPT) or Treatment A (in PT^3) the MAE is lower than in other cases. This result supports the idea that the reference point is (or is included in) the fixed lottery of the two involved in a binary choice as previously suggested (Stalmeier and Bezembinder 1999, Morrison 2000, Bleichrodt et al. 2001).

[26] We consider those respondents included in the two graphs at the middle of Figure 12.

Table 15. Best fit of CV relative values for individual responses and Mean Absolute Error for each approach

	F	W	X	V	S	R	N	L
Best fit (%)								
VAS	24.1	25.2	21.4	27.1	36.8	48.8	37.9	42.5
MSG_EUT	17.9	8.6	8.9	5.0	3.8	3.0	6.4	9.6
MSG_CPT								
$\boldsymbol{RP = U(\overline{w})}$	2.8	2.1	3.0	1.9	1.0	2.5	8.1	7.5
$\boldsymbol{RP = I(\overline{w})}$	1.7	0.7	0.9	0.5	1.5	2.7	6.4	8.9
$\boldsymbol{RP = I(\overline{w})} *$	14.2	17.1	17.5	20.3	16.3	11.4	12.8	12.3
$\boldsymbol{RP = D(\overline{w})}$	1.9	1.9	0.9	2.8	4.5	3.3	10.2	7.5
MSG_PT[3]								
$\boldsymbol{RP = Treat\ A}$	18.6	21.4	20.0	11.3	5.8	5.4	6.4	6.8
$\boldsymbol{RP = Treat\ A} *$	18.2	21.1	25.3	24.8	23.0	19.9	30.6	36.3
$\boldsymbol{RP = Treat\ B}$	2.4	3.6	2.1	6.4	7.5	5.2	13.6	15.1
Mean Absolute Error								
VAS	0.30	0.34	0.36	0.33	0.29	0.22	0.24	0.19
MSG_EUT	0.28	0.38	0.37	0.47	0.52	0.57	0.42	0.37
MSG_CPT:								
$\boldsymbol{RP = U(\overline{w})}$	0.28	0.37	0.36	0.45	0.51	0.56	0.43	0.39
$\boldsymbol{RP = I(\overline{w})}$	0.28	0.37	0.35	0.42	0.47	0.49	0.36	0.32
$\boldsymbol{RP = I(\overline{w})} *$	0.31	0.32	0.32	0.31	0.30	0.29	0.26	0.24
$\boldsymbol{RP = D(\overline{w})}$	0.28	0.36	0.35	0.42	0.48	0.51	0.39	0.36
MSG_PT[3]								
$\boldsymbol{RP = Treat\ A}$	0.28	0.36	0.34	0.41	0.47	0.48	0.36	0.32
$\boldsymbol{RP = Treat\ A} *$	0.32	0.31	0.31	0.30	0.30	0.30	0.26	0.23
$\boldsymbol{RP = Treat\ B}$	0.28	0.37	0.37	0.48	0.54	0.62	0.49	0.45

Note 1. $*$ denotes ratios with optimum parameters.
Note 2. *Best fit* percentages may not sum to 100 because some individual responses are best fit by more than one type of estimation.

8.2. Insensitivity in CV

In Jones-Lee et al. (1995) several explanations for the mismatch between the two methods are considered beyond the theory of decision under risk that respondents follow. In particular they discussed the insensitivity of CV responses to both, risk reduction and severity of the injury. They found that many subjects reported the same willingness to pay no matter the risk reduction size, either 4 or 12 in 100,000. This bias is not to arise in our study since all CV questions involve a risk reduction of 5 in 100,000. However in our study many respondents do not differentiate their responses for health states that are different in severity. In average, a respondent reports the same WTP in 4.7 out 10 pairs of health profiles. Its counterparts in MSG, average of pairs of health profiles for which subjects risk the same, is only 1.7.[27] Insensitivity is also evident if we analyse data on each NFRI pair. For example, 45%, of subjects that value V and X give the same response in CV, the corresponding percentage for MSG is 28.7%.[28] The insensitivity of CV is relatively higher when the difference in severity is higher. For instance, in the case of the pairs R-X, N-X and L-X the percentages are respectively 25.8% vs 14.9%, 28.4% vs 7.6%, and 25.2% vs 3.2%. Also a high proportion of subjects are willing to pay the same for the prevention of a risk of injury and for preventing a risk of death that goes from 25.8% to 57.1% for F and N respectively. This bias may derive in disproportionately high relative values for CV. In the *British Study* reasons for WTP insensitivity are related to "protest responses" that are either zero responses, or more than 500£ responses, or the same non-zero WTP response. However in our survey, respondents have the opportunity to answer high amounts of money (up to 300,000€ and more in case they ask for). Indeed zero responses are missing values when computing the CV relative values. Therefore only the non-zero responses can be possible as an explanation for the higher ratios estimated by CV with respect to MSG (Figures 7 to 11).

Theoretical or procedural explanations must be found in order to account for such differences in the two methods in order to improve design of future valuation questions and the interpretation of them. The consistency analysis here presented could serve as a guide for such a task. In the econometric analysis we find a significant effect of household income in probability of being strictly consistent (Table 18) and weakly consistent (Table 19) in CV responses. Since the coefficient in the latter is less significant and important this suggests that household income is mainly affecting *same responses* rather than changing the strict preference relation, for the health profiles pairs, between ranking and CV. If this is so it is reasonable to expect that the higher the income is the less the frequency of *same non-zero* WTP and then the lower the CV relative values are. We next analyse this argument by separating subjects according to the median household income (1,200€ a month) and restricting our sample to those respondent with ratios lower than or equal to unity (i.e. prefer an injury to death). For every NFRI the mean relative value is lower for respondents above median income. According to Mann-Whitney test for independent samples differences between the two income groups are significant for L (p-value<0.1), S (p-value <0.05), W (p-value <0.01) and F (p-value <0.01). For example, F mean relative value below median household income is 0.43 with 83 out of 309 same non-zero responses, while for the high income group is 0.29 with 41 out of 339 same non-zero responses. The corresponding figures for W are 0.45 vs 0.32, with 48, out of 199, and 25, out of 222, *same responses* respectively for below and above median household income. Moreover household income is found to have no significant impact on MSG relative values for any of the NFRIs but R (p-value<0.05) in the opposite direction, slightly lower ratio for low income group (0.27 vs 0.3).

We could think that insensitivity of CV responses is due to the embedding effect reported by Kahneman and Knetsch (1992) as if subjects are satisfied with spending some money on their own traffic safety without wondering the quality or quantity of the safety improvement.[29] We

---

[27] This information can be easily read in Table 8 by the next math operation: $WEAK_i^{CV(MSG)} - CONS_i^{CV(MSG)}$.

[28] This percentage is computed by subtracting the percentage in Table 12 to that of Table 11.

[29] Embedding has also been explained to some extent by *question misunderstanding* as reported by Fischhoff et al (1993). They found that respondents did not understand key details of CV questions accurately and if their answers were interpreted according to what subjects had understood, instead of the

could also think that subjects divide their budgets in different compartments each of these refers to the expenditure on goods or services as for example education, health care, housing, or road safety, so that the spending on preventing the risk of road injuries does not change with the severity of the accident. In any case we have to consider this insensitivity effect to be interacted with household income which in turn suggests that respondents do not change their willingness to pay when they should because they do not have enough financial resources to do so. Furthermore, given that even for those with above median household income the CV insensitivity is high, a straightforward interpretation is that spending on security devices is sufficiently small for a big part of the population compared to other expenses, like health services or food, such that few individuals are able to respond with considerable sensitivity within the limits of their budgets.[30]

8.3. Inconsistency in MSG

The MSG leads to slightly more inconsistency (less weak consistency), as can be seen in Tables 8 to 12, that could be interpreted as committed mistakes by respondents when reporting their preferences. These inconsistencies can be interpreted as changes in individual preferences as well, however we assume the change in preferences to affect inconsistency of CV and MSG to the same extent. Since we see a higher percentage of inconsistent responses in MSG we attribute this to mistakes. Regression analysis shows that those who do not understand probabilities commit more MSG inconsistencies (Table 19). It is relevant the fact that this factor has nothing to do with inconsistency in CV which can be explained because of the use of a unique risk reduction in all questions (respondent pays for a unique risk reduction from 15 to 10 in 100.000 for each NFRIs) so that a respondent just has to worry about the severity of the injury and not about probabilities. This finding is in line with the guidance of research suggested in Spackman et al. (2011), specifically it provides evidence about the nature of the imprecision of elicited preferences by both procedures so that, on the one hand, each method performance is assessed and, on the other hand, future survey design can be improved.

Another problem with this method has to do with the severest injuries N and L, ranked by a high percentage of respondents as worse than death. The design of MSG in this study does not allow this preference since responses to MSG are limited to the valuation of NFRIs better than or equal to death.[31] This implies that if a respondent's preferences are such that D$\succ$L then a weakly consistent response should arise in MSG (a strictly consistent response is not allowed) to be as coherent as possible. However, this is not the case in our study and a huge proportion of non weakly consistent responses show up. For example, about 75.6% and 78.1% of individuals who rank health states N and L respectively as worse than death have been inconsistent, showing the inversed preference relation, D$\prec$L or D$\prec$N. Nonetheless, the distortion of the average value of preventing an injury worse than death seem be important while it does not

---

actual questions then less embedding effect appeared. We think this is not the case in our study since respondents previously rank the NFRIs indicating their order of preferences (being aware of severity differences between them) and in spite of it they are willing to pay the same amount.

[30] The percentages of same non-zero responses for each of the rest NFRI and death by subjects with above median household income are: X (21.4%); V (28.4%); S (28.7%); R (40%); N (61.4%); L (66.1%). This is again considering only respondents with ratios lower than or equal to unity.

[31] The reason for this result is that MSG has a restricted range of responses on the level of risk assumed by individuals ranging from 0 to 1000 in 1000. With this range of responses expected utility of health states to be valued can never be worse than death utility. Specifically if $\boldsymbol{\pi = 1000}$ the respondent values the corresponding injury as equal to death, and if $\boldsymbol{\pi} < 1000$ then the injury is valued better than death.

seem to be relevant for median computation. The mean relative value for L is 0.92, that is at least 8% less than it should be, and the 25^{th}, 50^{th} and 75^{th} percentiles are 0.95, 0.998 and 0,999 respectively. The mean and percentiles figures for N are respectively 0.88, more than 12% below what it should be, 0.95, 0.980 and 0.999.

8.4. Future design

Results here exposed are interesting, although not totally solving the unmatched relative values by MSG and CV, and could be taken into account for future survey design and following research. Biases that have been detected in the two methods may not have any effect on the estimation of aggregate relative values, mean or median. For example, despite the occurrence of some inconsistencies in MSG for some subjects, this would have no effect in average if other respondents counterbalance by giving consistent responses in the opposite direction. Nonetheless, we believe that improving individual consistency and accuracy would not worsen estimations and some instructions can be given.

In first place an adapted version of MSG has to be used for the evaluation of NFRIs near or worse than death as in Patrick el al. (1994). An obvious prerequisite of the adapted MSG is usefulness for evaluating worse than death health profiles but also its framing has to be as similar as possible to that encounter for the valuation of equal or better than death injuries. It would be of interest to test whether this adapted MSG relative values of preventing NFRIs worse than death are more similar to their CV counterparts than in the case of health profiles better than death analysed in this study. This analysis could be of help for searching of theoretical foundations for CV-MSG mismatch.

There may be future studies in which only one questionnaire is provided, either MSG or CV. The econometric analysis may be used as a guide to assign the best procedure to each subject according to his/her characteristics. Of course, the fact that a person does not understand probabilities must be grounds for providing CV since it is an indicator of higher degree of inconsistent MSG responses. By contrast, a low income should be a reason for providing MSG because of the higher propensity of *same CV responses*, as exposed in section 8.2. An advance age is a reason for MSG because there are too many *same responses* in CV especially for those individuals older than 75 years old. For this age category the proportion of *same responses* in the valuation of each NFRI and death is larger than for the younger than 75 years old group. It becomes about double for W (51.4% vs 26.4%), X (52.2% vs 26.2%) and S (59.3% vs 34.1%).[32] Activities that have to do with the investment of time or money on changing health risk, like smoking and playing sports, or wealth risk, like gambling, appear to have a positive effect on strictly consistent CV responses suggesting that for these subgroups CV procedure could not be that bad. Eventually, those who drive are estimated to be more strictly consistent in CV (see Table 18) which in turn is a good reason for providing a CV questionnaire. A more subtle analysis tells us that this is due to more *same zero responses* by those who do not drive. For example 18.9% of non-drivers are willing to pay €0 for W and death, while the same percentage for drivers is 10.2%. This result could indicate that subjects interpret the safety device to be more valid when being a driver (i.e. using private transport modes). More emphasis must be

[32] Subjects aged more than 75 are also estimated to be less inconsistent in CV responses (more weakly consistent responses in Table 19). However, the average percentage of inconsistency decreases only in 4% in the valuation of each NFRI and death. Thus we maintain MSG to be a more suitable alternative for this group according to our results.

placed on alert respondents that the hypothetical safety improvement is independent of transport use.

If budget constraints appear to be associated with a high insensitivity of willingness to pay, then key details of CV questions could work to solve this problem. For example, in our study a series of payments for each CV question is displayed on the computer screen to facilitate WTP elicitation. The respondent must indicate if he would or would not pay such amount for sure. After that, the subject is asked to decide the amount he would pay, which should be placed between the highest amount of money he would pay and the lowest one he would not pay for sure. Although a rational decision maker should be neutral of the *displayed payment cards* it would be of interest to investigate if manipulations of these could derive in more/less consistency. Intuitively the use of smaller *displayed payments* could lead to lower WTP and thus less insensibility due to budget constraints.

Finally, correlation analysis shows that consistency in MSG is not independent of consistency in CV. In other words, there are factors affecting the probability of consistent responses in both methods. In this regard VAS scores are highly significant in predicting consistency in both procedures. The strength of preference expressed in the analogue scale is positively related to the probability of strict and weak consistency. On the one hand, this effect is so robust that it highlights the interval scale property of the VAS. However, on the other hand, this suggests that health states similar in value have more probability of not being valued consistently. Given this strong result it may be advantageous that respondents' have information about their previous valuations/responses so that he can resolve inconsistencies.

9. Conclusion

The relative value of preventing a Non-Fatal Injury is always larger when estimated by Contingent Valuation than with Modified Standard Gamble under both EUT and PT. In addition, differences between both methods worsen as severity of NFRIs to be valued decreases, which could be explained by insensitivity of CV responses. This result is consistent with previous work in the UK. Only the optimization of PT parameters could make MSG ratios similar to CV relative values. Nonetheless given that closer relative values have been computed using VAS scores we believe that this valuation process is very much like CV and research on the kind of cognitive process that VAS entails could give us some clues for explaining CV-MSG mismatch.

CV and MSG are well grounded in the two theories of decision under risk presented here. However, as it has been argued in the British Study, we think that enormous insensitivity of CV responses prevents it from being the best candidate for the valuation of NFRIs. For the MSG it has no problems like *embedding effect*. For example, in case that a subject has a low income, such that no higher than zero willingness to pay responses are reported in CV questions, MSG is useful for the estimation of the relative values. So, MSG is a more appropriate method because theoretically it provides the relative values that CV should provide in absent of restrictions, namely budgetary or any other type. Nonetheless, we have to keep in mind that we support MSG as long as there are difficulties with elicited preferences by CV.

Moreover, broad empirical evidence suggests that Prospect Theory outperforms Expected Utility Theory in description of decisions under risk. Presumably loss aversion and probability weighting are affecting subjects' decisions, so we consider PT relative values more suitable. Still there is a problem since PT relative values depend on the assumed reference point. There is no way of assessing the election of the RP within the scope of this study but assume that the RP is that which provide the best fit between CV and MSG. In this sense we have to say that Third Generation Prospect Theory provides the best fit and the reference point in MSG is Treatment A. This is consistent with the empirically finding that the reference point is encounter in the prospect (Treatment A), that remains constant, to be compared with another lottery (Treatment B) whose outcomes are varied by the respondent (Stalmeier and Bezembinder 1999, Morrison 2000, Bleichrodt et al. 2001).

Even more we could not say that MSG is the best method for respondents that consider a NFRI as worse than death. Many inconsistencies arise for the severest injuries N and L when valued by the version of MSG here presented. Therefore, while MSG is not adapted to be reliable in those cases we think that CV is a more reliable method.

Finally, our results indicate that "one best method" that suits every respondent could not exist. Individual characteristics could be taken into account in future surveys for providing each subject with a more appropriate questionnaire, either CV or MSG.

Appendices

Appendix 1. Theoretical Relative Value under CPT with alternative RP

In CV questions for eliciting WTP for reducing a risk of NFRI the *RP* is either $\{U, \overline{w}\}$ or $\{I, \overline{w}\}$ in prospect A. We now consider the case when RP=$\{I, \overline{w}\}$. Value of prospect A and B should equal:

$$I(\overline{w}) + W^{+}(1-\overline{q})[U(\overline{w}) - I(\overline{w})] = I(\overline{w}) + W^{+}(1-q)[U(w) - I(\overline{w})] - \lambda W^{-}(q)[I(\overline{w}) - I(w)]. \tag{33}$$

Now differentiating (33) with respect to q,

$$0 = -W^{+\prime}(1-q)[U(w) - I(\overline{w})] + W^{+}(1-q)\frac{\partial w}{\partial q}U'(w) - \lambda W^{-\prime}(q)[I(\overline{w}) - I(w)] + \lambda W^{-}(q)\frac{\partial w}{\partial q}I'(w);$$

rearranging and setting $q = \overline{q}$,

$$m_I = \left.\frac{\partial w}{\partial q}\right|_{q=\overline{q}} = \frac{W^{+\prime}(1-\overline{q})[U(\overline{w}) - I(\overline{w})]}{\lambda W^{-}(\overline{q})I'(\overline{w}) + W^{+}(1-\overline{q})U'(\overline{w})}. \tag{34}$$

Now we consider RP=$\{D, \overline{w}\}$ to compute theoretical m_D, this is willingness to pay for reducing the risk of death. Then the value of both prospect A and B are:

$$D(\overline{w}) + W^{+}(1-\overline{p})[U(\overline{w}) - D(\overline{w})] = D(\overline{w}) + W^{+}(1-p)[U(w) - D(\overline{w})] - \lambda W^{-}(p)[D(\overline{w}) - D(w)]. \tag{35}$$

Differentiating (35) with respect to p,

$$-W^{+\prime}(1-p)[U(w) - D(\overline{w})] + W^{+}(1-p)\frac{\partial w}{\partial p}U'(w) - \lambda W^{-\prime}(p)[D(\overline{w}) - D(w)] + \lambda W^{-}(p)\frac{\partial w}{\partial p}D'(w);$$

rearranging and setting $p = \overline{p}$,

$$m_D = \left.\frac{\partial w}{\partial p}\right|_{p=\overline{p}} = \frac{W^{+\prime}(1-\overline{p})[U(\overline{w}) - D(\overline{w})]}{\lambda W^{-}(\overline{p})D'(\overline{w}) + W^{+}(1-\overline{p})U'(\overline{w})}. \tag{36}$$

Now we divide (34) by (36) and obtain the relative value of preventing the risk of NFRI. Given that $\overline{p} = \overline{q}$ are very low probabilities (in CV questions $\overline{p} = \overline{q} = 0.00015$) we can assume the next expression holds to a good approximation:

$$\frac{m_i}{m_D} = \frac{U(\overline{w}) - I(\overline{w})}{U(\overline{w}) - D(\overline{w})} \times \frac{\lambda W^{-}(\overline{p})D'(\overline{w}) + W^{+}(1-\overline{p})U'(\overline{w})}{\lambda W^{-}(\overline{q})I'(\overline{w}) + W^{+}(1-\overline{q})U'(\overline{w})} \cong \frac{U(\overline{w}) - I(\overline{w})}{U(\overline{w}) - D(\overline{w})}. \tag{37}$$

Appendix 2. MSG Relative Values under CPT with alternative RP

If we consider $\{U, \overline{w}\}$ as *RP* then the value of *Treatment A* should equal to the value of *Treatment B* such that the next holds,

$$U(\overline{w}) - \lambda W^{-}(\theta)[U(\overline{w}) - D(\overline{w})] - \lambda[1 - W^{-}(\theta)][U(\overline{w}) - I(\overline{w})] = U(\overline{w}) - \lambda W^{-}(\pi^{*})[U(\overline{w}) - D(\overline{w})] \tag{38}$$

rearranging and scaling, $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$,

$$I(\overline{w}) = \frac{1 - W^{-}(\pi^{*})}{1 - W^{-}(\theta)}. \tag{39}$$

Now if we combine equations (11), scaling $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$, and (39) we obtain the next,

$$\frac{m_I}{m_D} = \frac{W^{-}(\pi^{*}) - W^{-}(\theta)}{1 - W^{-}(\theta)}. \tag{40}$$

If we consider $\{D, \overline{w}\}$ as *RP* then the value of *Treatment A* should equal to the value of *Treatment B*,

$$D(\overline{w}) + W^{+}(1 - \theta)[I(\overline{w}) - D(\overline{w})] = D(\overline{w}) + W^{+}(1 - \pi^{*})[U(\overline{w}) - D(\overline{w})], \tag{41}$$

rearranging and scaling, $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$,

$$I(\overline{w}) = \frac{W^{+}(1 - \pi^{*})}{W^{+}(1 - \theta)}. \tag{42}$$

Now if we combine equations (11), scaling $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$, and (42) we obtain,

$$\frac{m_I}{m_D} = \frac{W^{+}(1 - \theta) - W^{+}(1 - \pi^{*})}{W^{+}(1 - \theta)}. \tag{43}$$

Appendix 3. MSG Relative Values under PT[3] with $RP = Treat\ B$

Given that Treatment A and B are assumed to be independent, the states of the world in this case are the same as when the reference point is Treatment A. Hence the MSG answer π^{*} should be such that the value of both treatments are equal, this is $V[Treat\ A, Treat\ B] = V[Treat\ B, Treat\ B]$:

$$V[Treat\ B, Treat\ B] - \lambda W^{-}(\theta(1 - \pi^{*}))[U(\overline{w}) - D(\overline{w})] - \lambda\big[W^{-}(1 - \pi^{*}) - W^{-}\big(\theta(1 - \pi^{*})\big)\big][U(\overline{w}) - I(\overline{w})] + W^{+}((1 - \theta)\pi^{*})[I(\overline{w}) - D(\overline{w})] = V[Treat\ B, Treat\ B]. \tag{44}$$

rearranging and scaling $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$,

$$I(\overline{w}) = \frac{\lambda W^{-}(1-\pi^{*})}{W^{+}((1-\theta)\pi^{*})-\lambda W^{-}(\theta(1-\pi^{*}))+\lambda W^{-}(1-\pi^{*})}. \tag{45}$$

Now if we combine equations (18), scaling $U(\overline{w}) = 1$ and $D(\overline{w}) = 0$, and (45) we obtain the relative value of preventing a NFRI by MSG method under PT[3] with treatment B as reference,

$$\frac{m_I}{m_D} = \frac{W^{+}((1-\theta)\pi^{*})-\lambda W^{-}(1-\pi^{*})}{W^{+}((1-\theta)\pi^{*})-\lambda W^{-}(\theta(1-\pi^{*}))+\lambda W^{-}(1-\pi^{*})}. \tag{46}$$

Appendix 4. Procedure to adjust VAS scores to R-F effects

We want to compute relative values from (5) using the underlying values of injuries evaluated in the VAS, this is:

$$\frac{m_I}{m_D} = \frac{U(\overline{w})-I(\overline{w})}{U(\overline{w})-D(\overline{w})} = \frac{V_U-V_i}{V_U-V_D}. \tag{47}$$

where V_U, V_i and V_D are "context-free" or underlying values of normal health, the injury, and death respectively. However, instead of underlying values we have context-dependent VAS scores (VAS_i) which are determined by a weighted average of two values: the range value weighted with the factor ω, and; the frequency value weighted with the factor $1-\omega$ (see Wedell and Parducci, 1988):

$$VAS_i = \omega \times \frac{V_i-V_{min}}{V_{max}-V_{min}} + (1-\omega) \times \frac{rank_i-1}{N-1}. \tag{48}$$

where V_{max} and V_{min} are the maximum and minimum "context-free" values included in the evaluation context. And $rank_i$ is the rank of the injury within its context that goes from 1 to 6 for the worst and best health state respectively. Finally N=6 is the number of health states evaluated in the VAS exercise (four injuries plus normal health and death).

Now we can compute the range value like:

$$\frac{V_i-V_{min}}{V_{max}-V_{min}} = \frac{VAS_i}{\omega} - \frac{(1-\omega)}{\omega} \times \frac{rank_i-1}{N-1}. \tag{49}$$

Since we can compute the range value of each injury plus normal health and death we can calculate VAS relative values adjusted from R-F effects as:

$$\frac{m_I}{m_D} = \frac{\frac{V_U-V_{min}}{V_{max}-V_{min}} - \frac{V_i-V_{min}}{V_{max}-V_{min}}}{\frac{V_U-V_{min}}{V_{max}-V_{min}} - \frac{V_D-V_{min}}{V_{max}-V_{min}}} = \frac{V_U-V_i}{V_U-V_D}. \tag{50}$$

For the computation of range values in (49) it is necessary to know the value of ω. This parameter is usually estimated optimally for each data and is usually about 0.5. However we cannot perform this estimation because all the injuries evaluated in this study has the same rank in every context. Thus we will use $\omega = 0.46$ estimated by Robinson et al. (2001) for a VAS evaluation exercises of injuries analogous to those evaluated here.

Appendix 5. Tables

Table 16. Contingency tables for MSG and CV responses when valuing each pair of injuries. Row percentages of consistent respondents

F,W	NC^{CV}	C^{CV}	**F,S**	NC^{CV}	C^{CV}	**F,R**	NC^{CV}	C^{CV}	**F,N**	NC^{CV}	C^{CV}	**F,L**	NC^{CV}	C^{CV}
NC^{MSG}	*52*	48	NC^{MSG}	34	66	NC^{MSG}	38.9	61.1	NC^{MSG}	42.9	57.1	NC^{MSG}	53.9	46.2
C^{MSG}	46	54.1	C^{MSG}	26.9	73.1	C^{MSG}	24.2	75.8	C^{MSG}	24.8	75.2	C^{MSG}	23.4	76.7
W,S	NC^{CV}	C^{CV}	**W,R**	NC^{CV}	C^{CV}	**W,N**	NC^{CV}	C^{CV}	**W,L**	NC^{CV}	C^{CV}	**X,V**	NC^{CV}	C^{CV}
NC^{MSG}	42.2	57.8	NC^{MSG}	41.6	58.4	NC^{MSG}	40	60	NC^{MSG}	42.3	57.7	NC^{MSG}	51.1	48.9
C^{MSG}	32.1	67.9	C^{MSG}	28.7	71.3	C^{MSG}	29.8	70.2	C^{MSG}	25.7	74.3	C^{MSG}	44.3	55.7
X,S	NC^{CV}	C^{CV}	**X,R**	NC^{CV}	C^{CV}	**X,N**	NC^{CV}	C^{CV}	**X,L**	NC^{CV}	C^{CV}	**V,S**	NC^{CV}	C^{CV}
NC^{MSG}	44.6	55.4	NC^{MSG}	50.9	49.1	NC^{MSG}	53.9	46.2	NC^{MSG}	25	75	NC^{MSG}	61.3	38.7
C^{MSG}	27	73	C^{MSG}	20.5	79.5	C^{MSG}	27.2	72.8	C^{MSG}	26.9	73.1	C^{MSG}	45.3	54.7
V,R	NC^{CV}	C^{CV}	**V,N**	NC^{CV}	C^{CV}	**V,L**	NC^{CV}	C^{CV}	**S,R**	NC^{CV}	C^{CV}	**N,L**	NC^{CV}	C^{CV}
NC^{MSG}	63.2	36.8	NC^{MSG}	51.1	48.9	NC^{MSG}	26.1	73.9	NC^{MSG}	60.3	39.7	NC^{MSG}	66	34
C^{MSG}	31.7	68.3	C^{MSG}	34	66	C^{MSG}	33.5	66.5	C^{MSG}	46.1	53.9	C^{MSG}	52.8	47.2
M,F	NC^{CV}	C^{CV}	**M,W**	NC^{CV}	C^{CV}	**M,X**	NC^{CV}	C^{CV}	**M,V**	NC^{CV}	C^{CV}	**M,S**	NC^{CV}	C^{CV}
NC^{MSG}	100	0	NC^{MSG}	87.5	12.5	NC^{MSG}	100	0	NC^{MSG}	80	20	NC^{MSG}	77.3	22.7
C^{MSG}	27.4	72.6	C^{MSG}	31.4	68.6	C^{MSG}	27.1	72.9	C^{MSG}	37.3	62.7	C^{MSG}	44.1	56
M,R	NC^{CV}	C^{CV}	**M,N**	NC^{CV}	C^{CV}	**M,L**	NC^{CV}	C^{CV}						
NC^{MSG}	61.4	38.6	NC^{MSG}	63.4	36.6	NC^{MSG}	53	47						
C^{MSG}	52.7	47.3	C^{MSG}	67.1	32.9	C^{MSG}	70.2	29.8						

Note 1. It is shown NC^{CV} (*Non Consistent*) or C^{CV} (*Consistent*) when $C^{CV}=0$ or $C^{CV}=1$ respectively**.** The same apply to NC^{MSG} and C^{MSG}.

Table 17. Contingency tables for MSG and CV responses when valuing each pair of injuries. Row percentages of weakly consistent respondents

F,W	NWC^{CV}	WC^{CV}	**F,S**	NWC^{CV}	WC^{CV}	**F,R**	NWC^{CV}	WC^{CV}	**F,N**	NWC^{CV}	WC^{CV}	**F,L**	NWC^{CV}	WC^{CV}
NWC^{MSG}	0.0	100.0	NWC^{MSG}	11.7	88.2	NWC^{MSG}	15.8	84.2	NWC^{MSG}	22.2	77.9	NWC^{MSG}	14.3	85.7
WC^{MSG}	2.2	97.8	WC^{MSG}	0.4	99.6	WC^{MSG}	0.0	100.0	WC^{MSG}	0.4	99.6	WC^{MSG}	0.4	99.6
W,S	NWC^{CV}	WC^{CV}	**W,R**	NWC^{CV}	WC^{CV}	**W,N**	NWC^{CV}	WC^{CV}	**W,L**	NWC^{CV}	WC^{CV}	**X,V**	NWC^{CV}	WC^{CV}
NWC^{MSG}	10.0	90.0	NWC^{MSG}	16.7	83.3	NWC^{MSG}	11.1	88.9	NWC^{MSG}	0.0	100.0	NWC^{MSG}	5.1	94.9
WC^{MSG}	0.9	99.1	WC^{MSG}	0.0	100.0	WC^{MSG}	0.4	99.6	WC^{MSG}	0.0	100.0	WC^{MSG}	1.5	98.5
X,S	NWC^{CV}	WC^{CV}	**X,R**	NWC^{CV}	WC^{CV}	**X,N**	NWC^{CV}	WC^{CV}	**X,L**	NWC^{CV}	WC^{CV}	**V,S**	NWC^{CV}	WC^{CV}
NWC^{MSG}	13.3	86.7	NWC^{MSG}	18.7	81.2	NWC^{MSG}	28.6	71.4	NWC^{MSG}	25.0	75.0	NWC^{MSG}	7.7	92.3
WC^{MSG}	0.4	99.6	WC^{MSG}	0.0	100.0	WC^{MSG}	0.8	99.2	WC^{MSG}	0.8	99.2	WC^{MSG}	3.0	96.9
V,R	NWC^{CV}	WC^{CV}	**V,N**	NWC^{CV}	WC^{CV}	**V,L**	NWC^{CV}	WC^{CV}	**S,R**	NWC^{CV}	WC^{CV}	**N,L**	NWC^{CV}	WC^{CV}
NWC^{MSG}	6.9	93.0	NWC^{MSG}	9.1	90.9	NWC^{MSG}	16.7	83.3	NWC^{MSG}	3.6	96.4	NWC^{MSG}	5.0	95.0
WC^{MSG}	0.9	99.0	WC^{MSG}	1.3	98.7	WC^{MSG}	0.8	99.2	WC^{MSG}	2.6	97.4	WC^{MSG}	2.0	97.9
M,F	NWC^{CV}	WC^{CV}	**M,W**	NWC^{CV}	WC^{CV}	**M,X**	NWC^{CV}	WC^{CV}	**M,V**	NWC^{CV}	WC^{CV}	**M,S**	NWC^{CV}	WC^{CV}
NWC^{MSG}	72.7	27.3	NWC^{MSG}	85.7	14.3	NWC^{MSG}	100.0	0.0	NWC^{MSG}	71.4	28.6	NWC^{MSG}	25.0	75.0
WC^{MSG}	1.8	98.2	WC^{MSG}	3.0	96.9	WC^{MSG}	1.8	98.2	WC^{MSG}	4.5	95.5	WC^{MSG}	9.1	90.8
M,R	NWC^{CV}	WC^{CV}	**M,N**	NWC^{CV}	WC^{CV}	**M,L**	NWC^{CV}	WC^{CV}						
NWC^{MSG}	13.7	86.3	NWC^{MSG}	7.9	92.1	NWC^{MSG}	5.4	94.6						
WC^{MSG}	10.7	89.3	WC^{MSG}	8.8	91.2	WC^{MSG}	12.5	87.5						

Note 1. It is shown NC^{CV} or C^{CV} when $C^{CV} = 0$ or $C^{CV} = 1$ respectively. The same apply to NC^{MSG} and C^{MSG}.
Note 2. Row percentages shown.

Table 18. Effect of variables on Consistent Responses of CV and MSG. Regression Analysis

	CV				MSG			
	Random Effects			F. Effects	Random Effects			F. Effects
VARIABLES	**(1)**	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**
Age:								
18-30 (cons.)								
31-45	-0.279	-0.324	-0.407		-0.181*	-0.194**	-0.229*	
	(0.268)	(0.266)	(0.318)		(0.093)	(0.093)	(0.117)	
46-60	-0.514*	-0.545*	-0.562		-0.164	-0.172*	-0.170	
	(0.290)	(0.288)	(0.348)		(0.101)	(0.100)	(0.128)	
61-75	-0.624*	-0.570*	-0.598		-0.209*	-0.192	-0.204	
	(0.346)	(0.344)	(0.417)		(0.119)	(0.119)	(0.152)	
>75	-2.085***	-1.893***	-1.820***		-0.343**	-0.285*	-0.119	
	(0.477)	(0.475)	(0.567)		(0.161)	(0.160)	(0.204)	
Education:								
Less than second. (cons.)								
Secondary	0.396*	0.218	0.202		0.095	0.038	0.074	
	(0.236)	(0.239)	(0.284)		(0.082)	(0.083)	(0.104)	
Vocational	0.684*	0.526	0.376		0.042	-0.013	-0.104	
	(0.389)	(0.389)	(0.462)		(0.134)	(0.134)	(0.168)	
Tertiary	0.754***	0.414	0.386		0.165*	0.052	0.021	
	(0.286)	(0.301)	(0.357)		(0.099)	(0.104)	(0.130)	
Employment Status:								
Worker Priv. s. (cons.)								
Worker Pub. s.	-0.537	-0.632	-0.912*		-0.298**	-0.334**	-0.392**	
	(0.417)	(0.415)	(0.490)		(0.142)	(0.141)	(0.177)	
Self-employed	-0.253	-0.302	-0.411		-0.190*	-0.212*	-0.190	
	(0.337)	(0.336)	(0.396)		(0.115)	(0.114)	(0.142)	
Unemployed	-0.595**	-0.350	-0.401		0.007	0.077	0.075	
	(0.299)	(0.303)	(0.361)		(0.104)	(0.105)	(0.132)	
Inactive	0.033	0.098	0.355		-0.048	-0.030	-0.052	
	(0.270)	(0.268)	(0.326)		(0.093)	(0.093)	(0.119)	
Household income:								
0 – 1200€ (cons.)								
1200 – 1800€		0.695***	0.512*			0.193**	0.121	
		(0.226)	(0.274)			(0.078)	(0.100)	
> 1800€		0.917***	0.805**			0.311***	0.284**	
		(0.262)	(0.317)			(0.090)	(0.116)	
Smoker								
Non-smoker (cons)								
1-10 cig. a w.			0.648*				0.168	
			(0.339)				(0.125)	
11-20 cig. a w.			0.834**				0.117	
			(0.332)				(0.121)	
>20 cig. a w.			0.103				0.123	
			(0.422)				(0.153)	
Plays sport			0.466**				0.081	
			(0.224)				(0.082)	
Plays gambling			0.663***				0.143	
			(0.244)				(0.088)	

	CV				MSG			
	Random Effects			F. Effects	Random Effects			F. Effects
VARIABLES	**(1)**	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**
Happiness index			0.210*				0.040	
			(0.110)				(0.040)	
Driver			0.504**				-0.066	
			(0.250)				(0.091)	
Understands the risks			-0.808				0.324	
			(0.577)				(0.204)	
Absolute VAS score diff.								
VASdif			0.092***	0.086***			0.063***	0.065***
			(0.005)	(0.005)			(0.004)	(0.004)
VASdif^2			-0.001***	-0.001***			-0.0003***	-0.0003***
			(0.000)	(0.000)			(0.000)	(0.000)
Constant	0.759***	0.430	-2.931***		1.244***	1.146***	-0.802**	
	(0.286)	(0.297)	(0.868)		(0.099)	(0.103)	(0.313)	
Observations	10,029	10,029	10,015	6,218	10,040	10,040	10,026	9,226
Number of individuals	1,004	1,004	1,003	623	1,004	1,004	1,003	923

Note 1. Estimation coefficient above Standard Error, the latter shown in bracket.

Note 2. ***, **, and * mean coefficient is significant at 1%, 5% and 10% of error respectively.

Table 19. Effect of variables on Weakly Consistent responses of CV and MSG. Regression Analysis

	CV				MSG			
	Random Effects			F. Effects	Random Effects			F. Effects
VARIABLES	**(1)**	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**
Age:								
18-30 (cons.)								
31-45	0.120	0.122	0.224		-0.014	-0.019	0.027	
	(0.382)	(0.381)	(0.397)		(0.143)	(0.143)	(0.171)	
46-60	-0.029	-0.006	0.251		-0.083	-0.088	-0.012	
	(0.411)	(0.410)	(0.432)		(0.154)	(0.154)	(0.186)	
61-75	-0.487	-0.444	-0.100		0.042	0.044	0.207	
	(0.472)	(0.471)	(0.499)		(0.183)	(0.182)	(0.222)	
>75	1.522*	1.611*	1.971**		-0.096	-0.074	0.258	
	(0.833)	(0.833)	(0.845)		(0.244)	(0.245)	(0.295)	
Education:								
Less than secondary (cons.)								
Secondary	-0.543*	-0.658*	-0.614*		-0.017	-0.066	-0.012	
	(0.329)	(0.336)	(0.345)		(0.125)	(0.127)	(0.150)	
Vocational	-0.177	-0.264	-0.282		0.133	0.085	0.026	
	(0.557)	(0.562)	(0.578)		(0.212)	(0.213)	(0.252)	
Tertiary	-0.217	-0.478	-0.426		0.088	-0.011	-0.015	
	(0.411)	(0.436)	(0.448)		(0.154)	(0.161)	(0.191)	
Employment Status:								
Worker Priv. s. (cons.)								
Worker Pub. s.	-0.001	-0.065	-0.063		-0.024	-0.051	-0.109	
	(0.586)	(0.588)	(0.612)		(0.222)	(0.222)	(0.263)	
Self-employed	-0.154	-0.191	-0.191		0.054	0.021	0.075	
	(0.468)	(0.469)	(0.483)		(0.180)	(0.180)	(0.213)	
Unemployed	0.210	0.314	0.152		0.130	0.150	0.157	
	(0.440)	(0.449)	(0.461)		(0.162)	(0.164)	(0.195)	
Inactive	-0.051	-0.039	-0.001		-0.085	-0.085	-0.077	
	(0.380)	(0.379)	(0.400)		(0.142)	(0.141)	(0.172)	
Household income:								
0 – 1200€ (cons.)								
1200 – 1800€		0.049	0.241			-0.031	-0.139	
		(0.314)	(0.333)			(0.118)	(0.144)	
> 1800€		0.658*	0.727*			0.275**	0.236	
		(0.386)	(0.405)			(0.140)	(0.170)	
Smoker								
Non-smoker (cons)								
1-10 cig. a w.			-0.543				0.077	
			(0.385)				(0.181)	
11-20 cig. a w.			0.981**				0.372**	
			(0.468)				(0.183)	
>20 cig. a w.			0.676				-0.024	
			(0.571)				(0.224)	
Plays sport			0.205				0.188	
			(0.277)				(0.119)	
Plays gambling			-1.006***				0.007	
			(0.331)				(0.130)	
Happiness index			-0.192				-0.034	

	CV				MSG			
	Random Effects			F. Effects	Random Effects			F. Effects
VARIABLES	**(1)**	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**
			(0.137)				(0.059)	
Driver:			0.442				0.102	
			(0.300)				(0.132)	
Understands the risks			-0.970				0.607**	
			(0.821)				(0.281)	
Absolute VAS score diff.								
VASdif			0.045***	0.047***			0.077***	0.083***
			(0.009)	(0.009)			(0.006)	(0.007)
VASdif^2			-0.0004***	-0.0004***			-0.001***	-0.001***
			(0.000)	(0.000)			(0.000)	(0.000)
Constant	5.642***	5.530***	6.678***		2.581***	2.561***	0.611	
	(0.471)	(0.483)	(1.211)		(0.155)	(0.161)	(0.448)	
Observations	10,029	10,029	10,015	1,623	10,040	10,040	10,026	5,220
Number of individuals	1,004	1,004	1,003	163	1,004	1,004	1,003	522

Note 1. Estimation coefficient above Standard Error, the latter shown in bracket.
Note 2. ***, **, and * mean coefficient is significant at 1%, 5% and 10% of error respectively.

References

1. Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R. and Shuman, H. (1993). Report of the NOAA Panel on Contingent Valuation. *Federal Register,* 58 (10), 4601-4614.
2. Atkinson, A. B. (1995). *Income distribution in Europe and the United States*. Luxembourg Income Study.
3. Bleichrodt, H., Pinto, J. L., and Wakker, P. P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science, 47*(11), 1498-1514.
4. Bleichrodt, H., Abellan-Perpinan, J. M., Pinto-Prades, J. L., and Mendez-Martinez, I. (2007). Resolving inconsistencies in utility measurement under risk: Tests of generalizations of expected utility. *Management Science, 53*(3), 469-482.
5. Brooks R, Rabin R, de Charro F (eds). (2003). The Measurement and Valuation of Health Status. Using EQ-5D: A European Perspective. Kluwer: Dordrecht.
6. Cameron, A. C. (2005). Microeconometrics: Methods and applications / *A*. Cambridge; New York : Cambridge University Press.
7. Carthy, T., Chilton, S., Covey, D., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N., and Spencer, A. (1999). On the contingent valuation of safety and the safety of contingent valuation: Part 2 - the CV/SG "chained" approach. *Journal of Risk and Uncertainty, 17*(3), 187-213.
8. Chilton, S., Covey, J., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N., et al. (2002). Public perceptions of risk and preference-based values of safety. *Journal of Risk and Uncertainty, 25*(3), 211-232.
9. Fischhoff, B., Quadrel, M., Kamlet, M., Loewenstein, G., Dawes, R., and Fischbeck, P. (1993). Embedding effects - stimulus representation and response-mode. *Journal of Risk and Uncertainty, 6*(3), 211-234.
10. Gonzalez, R., and Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology, 38*(1), 129-166.
11. Greiner W. (2003). A European EQ-5D valuation set. In *The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective*, Chapter 8, Brooks R, Rabin R, de Charro F (eds). Kluwer: Dordrecht, 2003.
12. Harless, D. W., and Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica: Journal of the Econometric Society*, 1251-1289.
13. Hey, J. D., and Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society*, 1291-1326.
14. Jones-Lee, M., Loomes, G., and Philips, P. (1995). Valuing the prevention of nonfatal road injuries - contingent valuation vs standard gambles. *Oxford Economic Papers-New Series, 47*(4), 676.
15. Jones-Lee MW, Loomes G, Robinson A. (1994). What do visual analogue scales actually measure? Health Economists' Study Group (HESG) discussion paper, Newcastle Upon Tyne, UK.
16. Jones-Lee M. (1989). The economics of safety and physical risk. *Oxford: Basil Blackwell*.

17. Kahneman, D., and Knetsch, J. L. (1992). Valuing public-goods - the purchase of moral satisfaction. *Journal of Environmental Economics and Management, 22*(1), 57-70.

18. Kahneman, D., and Tversky, A. (1979). Prospect theory - analysis of decision under risk. *Econometrica, 47*(2), 263-291.

19. Morrison, G. C. (2000). The endowment effect and expected utility. *Scottish Journal of Political Economy, 47*(2), 183-197.

20. Neal, D., and Rosen, S. (2000). Theories of the Distribution of Earnings. Chapter 7 in *Handbook of income distribution*, *1*, 379-427. Atkinson, A. B., and Bourguignon, F. (Eds.). Elsevier.

21. Lichtenstein, S., and Slovic, P. (1971). Reversals of Preferences between Bids and Choices in Gambling Decisions. *Journal of Experimental Psychology* 89, 46–55.

22. Oliver, A. (2005). Testing the internal consistency of the lottery equivalents method using health outcomes. *Health Economics, 14*(2), 149-159.

23. Parducci, A. (1965). Category judgement: a Range-Frequency model. *Psychological Review,* 72(6), 407-418.

24. Parkin, D., and Devlin, N. (2006). Is there a case for using visual analogue scale valuations in cost-utility analysis? *Health Economics, 15*(7), 653-664.

25. Patrick, D. L., Starks, H. E., Cain, K. C., Uhlmann, R. F., and Pearlman, R. A. (1994). Measuring preferences for health states worse than death. *Medical Decision Making, 14*(1), 9-18.

26. Persson, Ulf, Trawén, Anna, Norinder, Anna, Hjalte, Krister, and Andersson, Henrik (2001). Relative Risk Values of Non-Fatal Traffic Injuries - A Comparison Between Contingent Valuation, Risk-Risk Trade Off and Standard Gamble Methods. *Swiss Journal of Economics And Statistics*, 137 (1). pp. 117-128.

27. Prelec, D. (1998). The probability weighting function. *Econometrica*, *66*(3), 497-527.

28. Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization, 3*(4), 323-343.

29. Robinson, A., Loomes, G., Jones-Lee, M. (2001). Visual analog scales, standard gambles, and relative risk aversion. *Medical Decision Making,* 21(1), 17–27.

30. Rottenstreich, Y., and Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, *12*(3), 185-190.

31. Schmidt, U., Starmer, C., and Sugden, R. (2008). Third-generation Prospect Theory. *Journal of Risk and Uncertainty, 36*(3): 203-223.

32. Schwartz, A. (1998). Rating scales in context. *Medical Decision Making*, 18, 236.

33. Spackman, M., Evans, A., Jones-Lee, M., Loomes, G., Holder, S., Webb, H., and Sugden, R. (2011). *Updating the VPF and VPIs: Phase 1: Final Report Department for Transport*. London: NERA Economic Consulting.

34. Stalmeier, P. F. M., and Bezembinder, T. G. G. (1999). The discrepancy between risky and riskless utilities: A matter of framing? *Medical Decision Making, 19*(4), 435-447.

35. Stalmeier P. F. M., Wakker P. P., and Bezembinder T. G. G. (1997). Preference reversals: violations of unidimensional procedure invariance. *Journal of Experimental Psychology*: *Human Perception and Performance* 23, 1196–1205.

36. Tversky, A., and Kahneman, D. (1992). Advances in prospect-theory - cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297-323.

37. Venkatachalam, L. (2004). The contingent valuation method: A review. *Environmental Impact Assessment Review, 24*(1), 89-124.

38. Viscusi, W. K., and Evans, W. N. (1990). Utility-functions that depend on health-status - estimates and economic implications. *The American Economic Review*, 80(3), 353–374.

39. Wedell, D. H., Parducci, A. (1988). The category effect in social judgment: experimental ratings of happiness. *Journal of Personality and Social Psychology*, 55(3), 341-356

Chapter 2. Valuation of Road Safety: Separate vs Joint Evaluation; Joint Evaluation context effects; and the reference-dependent evaluation of income

1. Introduction

Two major methods could be used to estimate Value of Statistical Life (VSL) and Value of Statistical Injury (VSI), either Revealed Preference (RP) or Stated Preference (SP). RP is based on actual microeconomic behaviour of individuals while SP relies on subjects' responses to a survey questionnaire. The former is widely used in the US and Canada and it is implemented in several works aimed at measuring the implicit trade-off between risk and wealth in several areas: decisions of workers in the labour market (see Kniesner et al., 2012; Viscusi, 2010; Kniesner and Viscusi, 2005; Viscusi and Aldy, 2003); additional expenditure in safety devices when buying a car (Andersson, 2005; Atkinson and Halvorson, 1990; Winston and Mannering, 1984); additional time spent on less risky activities like the use of pedestrians subways (Melinek, 1974), motorcycle helmet use (Blomquist et al. 1996), or seat belt use (Blomquist et al., 1996; Blomquist, 1979). However several drawbacks are found for this method: this approach requires assumptions about workers/consumers information and the choice set or on market failures; the use of proper econometric techniques is necessary to control for all the factors correlated with the involvement of people in risky activities; or the results are derive from preferences of a subset of the whole population (e.g. the labour force, car buyers, pedestrians, motorcycle riders).

On the contrary, SP procedures are based on surveys that contain hypothetical questions to elicit respondents' marginal rate of substitution of wealth for risk of death/injury (i.e. VSL/VSI). This method is widely used in European countries such as the UK (Jones-Lee et al., 1985; Jones Lee et al., 1995; Carthy et al., 1999), Sweden (Persson and Cedervall, 1991; Persson et al., 1995; Persson et al., 2001(a); Persson et al., 2001(b)), Spain (Abellán et al., 2011) and the Netherlands (De Blaeij et al. 2002). It is increasingly more common to use SP methods to evaluate the economic value of safety since several advantages appear for these procedures which make them a good alternative for the estimation of VSL and VSI. For example, this method has the property of representativeness; estimations are from national representative samples as recently pointed out by Viscusi and Huber (2012). Also, it is possible to control for the information that consumers face like for example the risk reduction that they pay for. Moreover, there are no other considerations by subjects than those affected by safety improvements, so no estimation bias arises on Willingness To Pay (WTP) responses due to preferences over other characteristics (e.g. a preference for a specific type of job or car) that are difficult to separate from the safety features. It also allows researchers to study different aspects of safety valuation such as: societal willingness to pay for preventing accident/fatalities depending on the person culpability that leads to a risk situation (Covey et al., 2010); people valuation of other's safety vs prevention of own risk (Hammitt and Haninger, 2010).

Eventually, the main concern about SP elicitation procedures is that preferences are estimated only from responses to hypothetical situations that may not adequately describe the actual behaviour of individuals in the marketplace. Also there is a concern on what kind of bias could systematically affect the elicitation of preferences through SP methods based on surveys (see Spackman, 2011). The understanding of why safety valuation varies among methods and contexts is a key issue in the road safety valuation literature and in economic analysis in general; at least for two reasons. In the first place the study of biases allows us to classify/appraise methods according to the extent that they are affected by systematic errors. In the second place, the study of biases could give us the route to control for them in order to elicit "bias-free" valuations. Moreover, the study of individual decisions in the "laboratory" is relevant to economic analysis because it could shed light on some behavioural patterns that can be taken into account by economic theory in order to improve descriptive models of agents' behaviour.

Two major SP methods are used for road safety valuation, namely Contingent Valuation (CV) and Standard Gamble (SG). We try to contribute to the understanding of these methods by analysing the results of a survey in which a sample of the Spanish population values Non-Fatal Road Injuries (NFRIs) and the prevention of a road fatality. The design of the survey allows us to examine several features of the valuation procedures. First we look at the effect of the Evaluation Mode (EM) which is either Joint Evaluation (JE), in which NFRIs are valued jointly with others injuries, or Separate Evaluation (SE), that entails valuation of each NFRI in isolation. Secondly, we analyse contexts effects within the JE mode; i.e. we explore the impact of the contextual stimuli (i.e. the contextual injuries) on the valuation of the target stimulus (the injury that is being valued). Finally, we consider the role of the reference income on the valuation of safety. Specifically, we try to disentangle the effect of the *past and expected future income* (the reference income) from the impact of current income.

In the survey analysed here respondents value NFRIs through CV questions and a modified version of Standard Gamble (MSG). Both methods are presented to respondents in a separate and joint mode. In the former case, only one injury is valued by the same individual so that no other injuries are presented during the valuation process. While in JE, several injuries are valued by the same individual and all the injuries are presented. In other research areas, previous works have demonstrated that valuation under these two EMs differs in several respects. In first place, preference reversals occur between the two modes. Hsee (1996) presented an experiment in which subjects reported their willingness to pay for a dictionary. Some subjects valued a dictionary A which had 10,000 entries and was like new. Other group of subjects valued dictionary B which had 20,000 entries and the cover was torn. It happened that subjects were willing to pay more for A than for B. Eventually, when individuals had the chance to value the two dictionaries at the same time they gave a higher valuation to B than to A. Also insensitive valuation responses are more probable to occur in SE as proposed by *General Evaluability Theory* (see GET by Hsee and Zhang, 2010). In Hsee et al. (1999) a higher sensitivity to attribute variability is found when the evaluation of a hypothetical applicant to a university is carried out in contexts more similar to JE. Specifically, in SE no statistical differences appeared between applicants evaluation even though they had varying attribute values given by their *Academic Potential Exam* scores. However when applicants' average score or min-max score range was reported before to the evaluation subjects valued significantly better those candidates with higher exam scores. More recently Chu et al. (2010) found that products from countries which are regarded as producers of high (low) quality are valued better (worse) when jointly presented with the same products from low-quality (high-quality) countries than when valued in a separate frame. Both, preference reversals between JE and SE, and insensitivity/sensitivity in SE/JE could be due to the *Evaluability Hypothesis* (Hsee, 1996; Hsee et al., 1999; and Hsee and Zhang, 2010) which is grounded in the idea that some difficult to evaluate attributes, hard to value in SE, become easier to assess in a JE mode because reference points (information about average alternative and/or attribute range) facilitate the task.

For the case of road safety we find that sensitivity to injury severity is higher in JE for both CV and MSG. However, value insensitivity encountered in SE is much more extreme for CV given that some injuries with varying severity are not evaluated statistically significantly different to each other's. While for MSG even in SE high value sensitivity remains because all injuries with varying severity are evaluated statistically significantly different to each other's. Therefore valuation of NFRIs seems to be less affected by the specific EM which is a factor to be considered for a suitable SP elicitation procedure. In addition, we find that a systematic effect of the EM is that those mild (serious) NFRIs are evaluated as more (less) severe in SE than in JE. This last result suggests that SE evaluation of milder (more severe) injuries should be downward-corrected (upward-corrected) to encounter JE.

Also our survey design allows us to test for contexts effects between different JE groups. Although those subjects that are in a JE mode evaluate the same number of injuries (four), some of these injuries vary in severity giving place to Parducci's range-frequency effects (see Parducci, 1965; or Wedell and Parducci, 1988 for a description of these effects in the social judgment domain). These effects have already been detected in the injury valuation domain for the Visual Analogue Scale (VAS) (see Robinson et al., 2001). Therefore we can test whether this kind of effects can also be encounter in CV and MSG responses that theoretically are based on preferences rather than on categorical classification or rating values. Range Frequency Theory (RFT, Parducci, 1965) is a model that predicts contrast effects. In this case it is expected that an injury is valued as more severe (less severe) when it is evaluated in a context jointly with other mild (serious) injuries. In addition, we analyse another possibility called *anchor effects*. This is a stylized heuristics that consists on the empirical finding that judgmental responses given by individuals are affected by initially presented injuries in such a way that the eventual valuation is biased toward the prior stimulus. In our survey, the evaluation of the same injuries in different JE groups could be preceded by other injuries with varying severity so that it could lead to anchoring. The effect of initial values is early described by Lichtenstein and Slovic (1971) and also found by Tversky and Kahneman (1974). For example, Tversky and Kahneman carried out an experiment where subjects were asked to estimate the percentage of African countries in the United Nations. Previously to their answer a random number determined by spinning a wheel of fortune was presented. As a consequence, responses varied with the initially value presented. The median estimates were 25% and 45% for groups that received numbers 10 and 65, respectively, as starting points. Furnham and Boo (2011) review the anchoring literature emphasizing that this is a robust and pervasive effect presented in different domains including: general knowledge questions (Epley and Gilovich, 2001); probability estimates (Chapman and Johnson, 1999); legal judgments (Englich and Soder, 2009), valuations and purchasing decisions (Ariely et al., 2003), forecasting (Critcher and Gilovich, 2008), negotiation (Galinsky and Mussweiler, 2001) and self-efficacy (Cervone and Peake, 1986).

Our results indicate that context effects between different JE groups do affect more to CV responses than to MSG valuations. So not only MSG seems to be less affected by JE-SE modes but even within JE it is again more consistent between different contexts than CV. Moreover predicted effects by RFT do not seem to explain the differences found between JE groups. Neither anchoring does provide a good account of context effects. We rather find that some differences between contexts are consistent with RFT and other differences are better explained by anchoring which makes it difficult to predict a systematic bias.

Eventually we analyse the effect of the reference income (past and expected future income). In our survey we ask respondents to report their current monthly income and *their normal monthly income once they consider various stages of low/high earnings throughout their entire lives*. The latter can be referred to as "permanent" or "normal" income. Accordingly, we generate three income frames: a frame of losses (those with current income below their normal income); a frame of gains (subjects with current income higher than normal income); and a neutral frame (with current income equal to normal income). Our main interest is in the effect of income scenarios on safety valuation. However, we first investigate whether those scenarios affect happiness/life satisfaction in a similar way that it has been shown in previous studies of subjective wellbeing.

We find that those in a gain frame report a higher happiness (through six life satisfaction questions) than those in a neutral frame, and than those in a loss frame. Given that this occur even controlling for current income, those in a loss frame presumably has a higher financial capacity (a higher past and expected future income) than those in a neutral/loss scenario. We think that a plausible explanation is that respondents use their permanent income as a reference point with respect to which they evaluate their current income rather than as a measure of their financial capacity. This is also the best explanation given that we control in the econometric analysis for other determinants of happiness as age, gender, self-reported health, marital status, minor children, dependent elderly at home, employment status and education. This result is consistent with the fact that happiness depends on relative income rather than on absolute income (see Easterlin 1974, 1995 and 2001). However, given our analysis we cannot reject an absolute income component affecting happiness (Kahneman, 2008; Stevenson and Wolfers, 2008; or Clark et al., 2008) since we find that those subjects with high (low) current and permanent income are happier (less happy). Hsee et al. (2009) find that that both relative and absolute consumption of some goods influence happiness.

The rest of our analysis is aimed at exploring if this framing effect on reported happiness is also relevant for decisions that should be based on utility. For this purpose we test if WTP and Willingness To Risk (WTR) responses to CV and MSG questions, respectively, are different for those three different frames. We find that WTP for avoiding a risk of a road fatality is higher for those in a loss/gain frame with respect to those respondents in a neutral situation. We also find that for the majority of the injuries evaluated the WTP is higher in the gain frame with respect to both the neutral and loss frame, while no differences between those two latter exist. We show that this result can be explained by a general reference dependent utility function of income with the typical properties of loss aversion and diminishing sensitivity (Kahneman and Tversky, 1979; and Tversky and Kahneman, 1992). In contrast to these findings in the case of CV, we do not encounter framing effects in MSG responses for the majority of the injuries. Given that theoretical MSG responses under a reference dependent utility function do not dependent on changes in the marginal utility of income, we think that framing effects in CV are driven mainly by changes in the marginal utility of income between different frames: specifically the marginal utility is lower in the gain frame than in the loss frame (as loss aversion suggests) and higher in the loss frame than in the neutral point (as convexity in the loss domain suggests).[1]

[1] We will see in the theoretical analysis that the left partial derivative of income is what really matters.

In the next section, survey details are exposed. After that, section 3 will describe the two state preference methods. Then results for the SE-JE analysis are exposed in section 4. After that the results for JE context effects is considered in section 5. In section 6 we perform the analysis of the effect of reference-dependent income on safety valuation. Eventually, section 7 concludes.

2. The survey design

The survey here analysed is the same survey previously studied in Chapter 1 of this thesis. In the present work we further examine individual responses to test new hypotheses. This is possible because some features were incorporated to the design in such a way that subjects were assigned to different groups that vary with respect to the evaluation mode and the contextual stimuli. Comparison of these groups makes possible the examination of the evaluation mode and context effects. In addition the consideration of new survey information makes possible the analysis of the current and the permanent (the reference) income. In this section we display the basic characteristics of the survey focusing on the specific features that allow the present study.

The Spanish Road Traffic Directorate General (DGT) funded a research project with the aim of estimating road safety valuation by a nationally representative sample of the Spanish population. The survey is conducted by interviewers hired by the project's researchers during the first half of 2011 through interviews taken place in the home of the respondents with the help of a laptop where all the questions are illustrated by a computer program. A set of 55 questions divided into four parts are presented to a Spanish nationally representative sample of 2,016 individuals. The first part of the questionnaire collects information about the use of road transport by respondents and some comprehension questions. Secondly, respondents rank the NFRIs they are going to value and place them in a Visual Analogue Scale (VAS) ranging from 0 to 100. The core of the questionnaire is comprised by the Modified Standard Gamble (MSG) and Contingent Valuation (CV) questions through which respondents value Health States (HSs).[2] For the rest of the questionnaire is related to the collection of socio-demographic information.

Eight different NFRIs are used for valuation. These are the same as those shown in Figure 1 in Chapter 1 of this thesis.[3] These are analogous to those injuries valued in Jones-Lee et at. (1995). Each HS presents different level of seriousness in some attributes like time in hospital, the extent and duration of pain, degree and length of restrictions to leisure and work activities, degree of physical and mental ability, and independency for basic physical needs. So NFRIs extend over a wide range from the milder ones, like F or W, to the most serious, like N or L. It can be considered that there is an objective preference order with respect to the severity of the NFRIs used in the survey. This idea is supported by the fact that it can be assumed that people's preferences over the seriousness of health attributes are monotone and that the eight NFRIs are increasing in the severity of the attributes. Accordingly, it may be considered an objective preference order as $F \succcurlyeq W \succcurlyeq X \succcurlyeq V \succcurlyeq S \succcurlyeq R \succcurlyeq N \succcurlyeq L$. Given the varying severity of the health states we expect the valuation to switch. As a consequence one of the main analyses we

[2] In what follows the damage after an accident, prevention of which is being valued, shall be appointed either as health profile, health condition, health state (HS) or NFRI.

[3] All the interviews are carried out in Spanish though in the present article it will be shown the English translation of the information that respondents have to deal with.

perform is the comparison of the sensitivity of valuation to the injury severity in both separate and joint evaluation.

Eight questionnaires are presented to different groups of respondents. In Table 1 it is shown the differences between these eight groups that are due to the specific NFRIs valued. All the groups face MSG questions to value the prevention of one HS in Separate Evaluation and four HSs in Joint Evaluation. With respect to CV questions, groups 1 to 4 face the valuation of preventing one separate NFRI while groups 5 to 8 respond to the valuation of preventing four different injuries in JE. Also all the groups value in CV the prevention of a risk of road fatality. The order of the evaluation is that of the columns in Table 1. After an introductory questionnaire respondents value one NFRI in Separate Evaluation with MSG (column 3). Then they value the reduction of risk of a road fatality in SE with CV (column 4). After that, individuals value the same previous NFRI jointly with three additional HSs (column 5). Eventually, CV questions are carried out in SE, for groups 1 to 4, and JE, for groups 5 to 8, to evaluate one and four NFRIs respectively (column 6).

With this survey design we are able to analyse the effect of the EM in both MSG and CV by comparing the evaluation of NFRIs in JE and SE. With respect to MSG questions each of the eight injuries is evaluated in SE and JE. For example, valuation of injury F by group 1 in SE can be compared to the valuation of the same injury by the same group in JE. This is called a "within group" comparison. Also valuation of injury F by group 1 in SE can be compared to the valuation of the same health state by groups 1, 4, 5 and 8 in JE. This is referred to as a "between groups" comparison. Notice that for all the injuries we can study within and between comparisons of SE vs JE in the case of MSG. In CV only four health conditions are valued in isolation: F, X, V and R. Moreover only between groups comparisons are possible in the case of CV. For example, valuation of injury F by group 1 can be compared to valuation of the same injury by groups 5 and 8.

Each joint evaluation setting is characterised by four NFRIs for valuation. These NFRIs are different for each group which allows us to explore context effects for both CV and MSG by comparing these different contexts. The mains hypothesis in this case is weather the valuation of a specific health state depends on the contextual stimuli (other health states) that are being valued jointly in the same elicitation process. In this sense four distinct contexts can be consider in the case of MSG that are characterised by the next NFRIs: Groups 1 and 5 with F, W, N and L; Groups 2 and 6 with X, V, N and L; Groups 3 and 7 with X, V, S and R, and; Groups 4 and 8 with F, W, S and R. The effect that we expect is driven by the fact that each injury is evaluated in two different contexts. For example, in groups 1 and 5 F and W is evaluated jointly with, N and L. While in groups 4 and 8 they are presented jointly with S and R. Given that S and R are two less serious health conditions than N and L this could affect the valuation of F and W across contexts. In the case of CV the contexts are the same once we consider only groups from 5 to 8.

As can be noticed only CV valuation of preventing a risk of a fatality is possible. Given the nature of MSG this method cannot be used for this task. The reason for incorporating in the survey the evaluation of risk of death by CV is to be able to calculate the relative value of preventing a risk of injury with respect to a risk of fatality. This was the main interest in Chapter 1 of this thesis. However, in the present study it is not this computation that we are interested in. Also, notice that the risk of a fatality has been evaluated only in SE and therefore we cannot test for EM effects or JE context effects. However our main interest is the use of responses to CV valuation of risk of death to analyse the effect of the reference income (permanent income). Nonetheless, this effect is also considered on the evaluation of NFRIs by CV and MSG.

Table 1. Sample size and Non-Fatal Road Injuries valued by group

Group	Observations	MSG (SE)	CV (SE)	MSG (JE)	CV (SE/ JE)
1	254	F	D	F, W, N, L	F
2	251	V	D	X, V, N, L	V
3	256	X	D	X, V, S, R	X
4	251	R	D	F, W, S, R	R
5	253	L	D	F, W, N, L	F, W, N, L
6	250	N	D	X, V, N, L	X, V, N, L
7	248	S	D	X, V, S, R	X, V, S, R
8	253	W	D	F, W, S, R	F, W, S, R
Total	**2,016**				

3. The preference elicitation methods

3.1. Contingent valuation

This method involves directly asking respondents about the amount of money they are willing to pay for reducing the risk of traffic accident. An example of a formulated CV question in the survey is:

> *Suppose your risk of injury such as W as a result of a traffic accident is 15 in 100,000 and that there exists a safety device that will reduce your risk of health status such as W in a traffic accident in 5 / 100,000, from 15 in 100,000 to 10 in 100,000.*

Respondents are told that the safety device is for single use. That way we are able to individualize the value of preventing a non-fatal accident and avoid responses taking into account other's people safety. For example, a head of family would think that if (s)he travels often with other household members they also benefit from this safety device and thus be willing to pay higher amounts of money. This feature is crucial when aggregating willingness to pay responses in order to not overstate the value of risk reduction. Respondents are told that this device works in all transport modes in order to avoid different responses according to their transport habits and make them believe that they would benefit from the safety increment anyway. Another aspect that deserves comment is the fact that respondents are explicitly told that the safety device has one year of duration. This latter characteristic of the CV questions is important because it implies that assessment of traffic safety programs has to be made according to how many non-fatal accidents can be avoided in one year and how much society is willing to pay for that. All CV questions, for the valuation of prevention of NFRIs, are made assuming that the respondent pays for a risk reduction of 5 in 100,000 accidents (from 15 in 100,000 to 10 in 100,000). For more information about the CV procedure followed in our survey see section 5.2 in Chapter 1 of this thesis.

Given CV responses the interesting computation for policy purposes is the Marginal Rate of Substitution (MRS) of wealth for risk of a non-fatal accident. That is the amount of money a person is willing to give up from her wealth, w, for an infinitesimal reduction of the probability of injury, q. Given that the safety improvement assumed in CV question is sufficiently small we can compute the MRS as the ratio between the amount of money a respondent is willing to pay for the safety improvement, wtp_I, and the risk reduction considered:

$$m_I = \left.\frac{\partial w}{\partial q}\right| \cong \frac{wtp_I \ (euros)}{risk\ reduction} \quad (1)$$

Although it is the MRS the figure that has to be used in public programs assessments we will show absolute willingness to pay in our results because our aim is to analyse row CV responses in different group of respondents.

3.2. Modified Standard Gamble

With the modified standard gamble method health states valuation is done by individuals by making choices between two hypothetical risk situations. One situation, denoted by *Treatment A*, is such that the individual suffers a particular non-fatal injury, which is the one being valued, with the probability $(1-\theta)$ and otherwise (s)he dies, with $\theta > 0$ probability. Another situation, denoted by *Treatment B*, is such that the individual continues with his/her normal health with probability $(1-\pi)$ and otherwise he dies, with $\pi > 0$ probability. The objective is to find a π^* such that the individual is indifferent between this two situations given a level of θ. In the study concerned here the parameter θ is fixed for every choice so that the respondent makes repeating choices depending on different levels of π that are suggested in a way such that the conclusion of the valuation of the non-fatal injury implies to obtain the indifference level.

In our survey θ is equal to 0.001 (1 in 1000). An example of a formulated MSG question in the survey is:

> *Suppose that you had a traffic accident and that, in case of not receiving medical care, you could die. There exists two treatments that, in principle, could be applied to your case: Treatment A and Treatment B. Suppose that with treatment A 999 of 1000 people have state V, while 1 in 1000 treated people dies. With treatment B the chances of dying are 400 in 1000 and the chances of returning to their normal health before the accident are 600 in 1000.*

Notice that $\pi = 0.4$ in this example and two possible responses arise. On the one hand, an individual can choose treatment A (treatment B) giving place to another question with $\pi < 0.4$ ($\pi > 0.4$) in order to get closer to the estimation of risk of death in treatment B that makes the individual be indifferent between both situations. On the other hand, the individual can report that both treatments are equally preferred implying that the indifferent risk of death level is 400 in 1000 ($\pi^* = 0.4$).[4] Hence MSG responses could vary from 1 in 1000 to 1000 in 1000. For more details about the MSG procedure see section 5.1 in Chapter 1 of this thesis.

[4] In the next, *Treatment B* will be referred to as the alternative lottery that gives a probability of dying π and a probability of resulting in normal health state equals to $1-\pi$.

Eventually MSG responses should be used to compute relative monetary value of avoiding a NFRI with respect to the value of avoiding a statistical life as in Carthy et al. (1999) in order to compute a benchmark for policy decision making. In this chapter we are mainly interested in analysing the change of MSG responses themselves in different contexts in order to understand how this benchmark can be affected.

3.3. The framing of the Separate and Joint Evaluation

Figure 1 is a screenshot of the CV question for group 1 in Separate Evaluation. First of all, subjects are explained some instructions and shown the description of the only injury they are going to value, in this case injury F (left panel of Figure 1). After that the evaluation screen is displayed (right panel of Figure 1) and the elicitation of the Willingness to Pay begins. During the whole process only the description the injury being valued is shown to the respondents in a paper. The separate evaluation is performed in a similar way for the MSG case.

We try to put subjects in a Joint Evaluation mode using two main changes in relation to the single question. First, using some similar features to Advance Disclosure (Bateman et al, 2007) subjects are shown a first screen with the four health states that they have to evaluate. They are explained that road traffic accidents could generate health problems of different severity and they are shown the four health states in the screen, as the example in Figure 2 for the case of CV in group 5. They are "announced" that are going to be asked four different willingness to pay questions (or MSG questions) for preventing each of the four health problems. Then they move to a different screen where they are asked exactly the same WTP or MSG question that is asked in the Separate Evaluation Mode, although for each of the four injuries. However, there is another difference between Joint and Separate at this point, namely, interviewers are instructed to have all four cards (in paper) in front of the subject in every case. In this way, they can compare the severity of the four injuries when responding to the WTP or MSG questions. The location of each description and the order of the evaluation of the NFRIs are randomized.

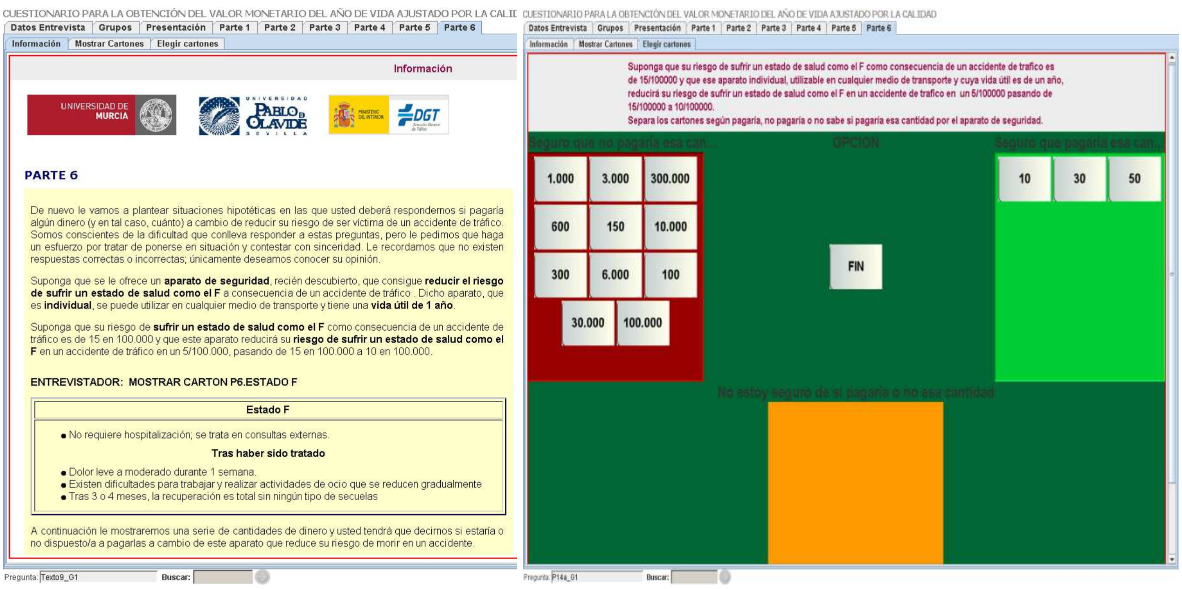
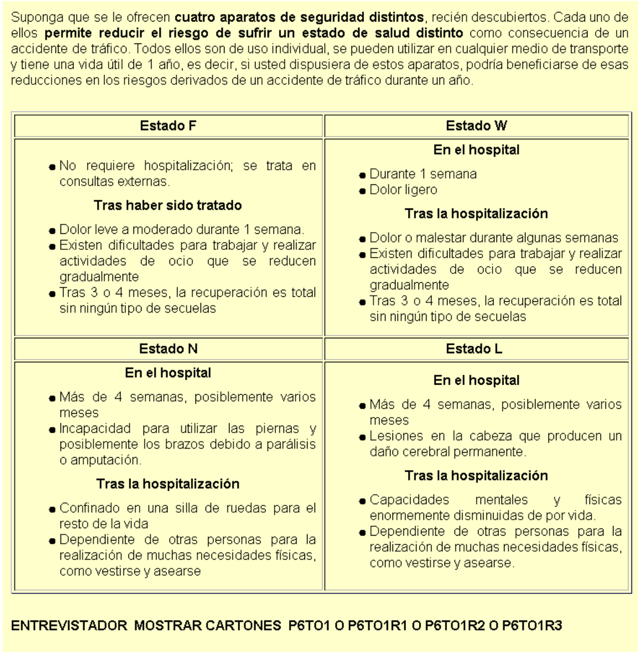


Figure 1. Screenshot for CV question in Separate Evaluation. Group 1



Suponga que se le ofrecen **cuatro aparatos de seguridad distintos**, recién descubiertos. Cada uno de ellos **permite reducir el riesgo de sufrir un estado de salud distinto** como consecuencia de un accidente de tráfico. Todos ellos son de uso individual, se pueden utilizar en cualquier medio de transporte y tiene una vida útil de 1 año, es decir, si usted dispusiera de estos aparatos, podría beneficiarse de esas reducciones en los riesgos derivados de un accidente de tráfico durante un año.

Estado F	Estado W
• No requiere hospitalización; se trata en consultas externas. **Tras haber sido tratado** • Dolor leve a moderado durante 1 semana. • Existen dificultades para trabajar y realizar actividades de ocio que se reducen gradualmente • Tras 3 o 4 meses, la recuperación es total sin ningún tipo de secuelas	**En el hospital** • Durante 1 semana • Dolor ligero **Tras la hospitalización** • Dolor o malestar durante algunas semanas • Existen dificultades para trabajar y realizar actividades de ocio que se reducen gradualmente • Tras 3 o 4 meses, la recuperación es total sin ningún tipo de secuelas
Estado N	**Estado L**
En el hospital • Más de 4 semanas, posiblemente varios meses • Incapacidad para utilizar las piernas y posiblemente los brazos debido a parálisis o amputación. **Tras la hospitalización** • Confinado en una silla de ruedas para el resto de la vida • Dependiente de otras personas para la realización de muchas necesidades físicas, como vestirse y asearse	**En el hospital** • Más de 4 semanas, posiblemente varios meses • Lesiones en la cabeza que producen un daño cerebral permanente. **Tras la hospitalización** • Capacidades mentales y físicas enormemente disminuidas de por vida. • Dependiente de otras personas para la realización de muchas necesidades físicas, como vestirse y asearse.

ENTREVISTADOR MOSTRAR CARTONES P6TO1 O P6TO1R1 O P6TO1R2 O P6TO1R3

Figure 2. Screenshot for CV question in Joint Evaluation. Group 5

4. Joint vs Separate Evaluation

4.1. Testing value sensitivity in SE and JE

Value sensitivity in the context of this study is defined as the extent to which individuals respond differently to different NFRIs with varying severity. According to value sensitivity it is expected that respondents are more willing to risk their lives when avoiding a severe NFRI like L than when avoiding a mild injury like F. Also it is expected that the amount of money that respondents are willing to pay for avoiding the former injury is higher than for avoiding the latter. In short, what is expected is that MSG and CV responses change with the severity of the health profiles.

In the analysis performed below there are two hypotheses to test: first, whether there is value sensitivity in SE or in JE (*H1*); and second, whether value sensitivity is higher/lower in SE with respect to JE (*H2*).

Both issues can be addressed by analysing graphically the responses in JE and SE. However, it is possible to test formally these hypotheses as follow. The first hypothesis is tested by performing non-parametric equality of distribution test of valuation of NFRIs. If MSG or CV responses for the evaluation of two NFRIs (for example W and X, or S and R) are not statistically significantly different then value insensitivity appears between this two injuries.

The second hypothesis is not directly testable given the nature of SE mode (each individual values only one NFRI in the separate frame). However, it is possible to count the number of HSs pairs that has been evaluated differently in each EM and used that as a measure of value sensitivity. Also we test if the same NFRI evaluated in SE is equally valued in JE. This analysis makes it possible to conclude if the responses range, the distance between the response for the mildest and the severest injuries, is higher in SE or JE. For example, if the evaluation of F is not significantly different in JE and SE, and the responses for L statistically do not change in both evaluation modes then we can conclude that the value sensitivity in both EMs is the same or, in other words, that the response range is the same. The non-parametric test of equality of distribution in SE and JE for each NFRI will conclude about the relative value sensitivity of both modes.

The non parametric tests performed below are the Wilcoxon signrank (W-SR) and ranksum tests (W-RS). The former is for testing the equality of distribution when the two variables to compare are from the same group (i.e. within group comparisons). For the latter, is performed in case that the variables to compare are from different subsamples (i.e. between groups analysis). We will say that the two variables that we are comparing do not have the same distribution if we can reject the null hypothesis at 10%.

4.2. Contingent Valuation

In Figure 3 mean and median CV responses are shown for each of the NFRI that has been evaluated in both contexts SE and JE. A positive slope of the graph lines is expected, this is respondents pay more to avoid more severe injuries. Also means and medians distance between evaluation responses of adjacent HSs is shown.

First, the response range is wider for JE. In SE mean willingness to pay goes from €210 (for F) to €577 (for R) while the same range in JE is from €80 to €764. Second the distance between adjacent NFRIs is always higher in JE. Wilcoxon tests conclude that JE is a value sensitive EM since it is rejected the null hypothesis of equality of distribution between adjacent NFRIs (p-value<0.1). The same does not apply to SE since the distribution of responses to F and X is not statistically significantly different (p-value>0.1) and the median response is the same; €50 for both injuries. Neither it is possible to reject the null hypothesis of equality of distribution between evaluation of V and R (p-value>0.1) that again have the same median (€100). Hence, no value sensitivity is found between the valuation of the milder pair of injuries and the more serious pair. In SE only value sensitivity between X and V is found (p-value<0.1).

In conclusion, with respect to H1 we find that JE is a value sensible mode for the evaluation of all the NFRI we have in our survey for CV questions. On the other hand, SE is not value sensible in the evaluation of the two pairs F-X and V-R. With respect to H2 we find that the value sensitivity in the evaluation of the pair X-V is the same for SE and JE.

One derivation of GET is that the value or utility function is less linear in SE than in JE (see Fig. 3 in Hsee and Zhang, 2010). The reason for this prediction is that people in SE do not differentiate much between incremental difference of an attribute (e.g. in the case of money between €10 and €20) while they are more value sensitive when the difference is categorical (e.g. between €0, no money, and €10, some money). In our framework we can think that there is a categorical difference between being in normal health and suffer any of the injuries, while the difference between injuries is mainly incremental (i.e. more or less pain, recovery time, the extent of work/leisure limitations). We show the utility function derived from CV responses in JE and SE in Figure 4 including the four NFRIs with normal health, U, and death, D.[5] As can be

[5] It can be shown that the ratio of MRS of wealth for risk of injury and death is related to the utility of wealth in normal health, injury or death state, as $\frac{m_I}{m_D} = \frac{U(w)-I(w)}{U(w)-D(w)}$. This is shown in section 2 of Carthy et al. (1999) and in Chapter 1 of this doctoral thesis it is demonstrated that this also holds, to a good approximation, when there is probability weighting and loss aversion. Scaling $U(w) = 1$ and $D(w) = 0$

seen the utility function in SE is less linear than in JE. In SE respondents differentiate a lot, in terms of utility, between normal health, U, and injury F, while the incremental difference between F, W and V is not very important. Notice that this pattern is not seen in JE or at least is far less pronounced.

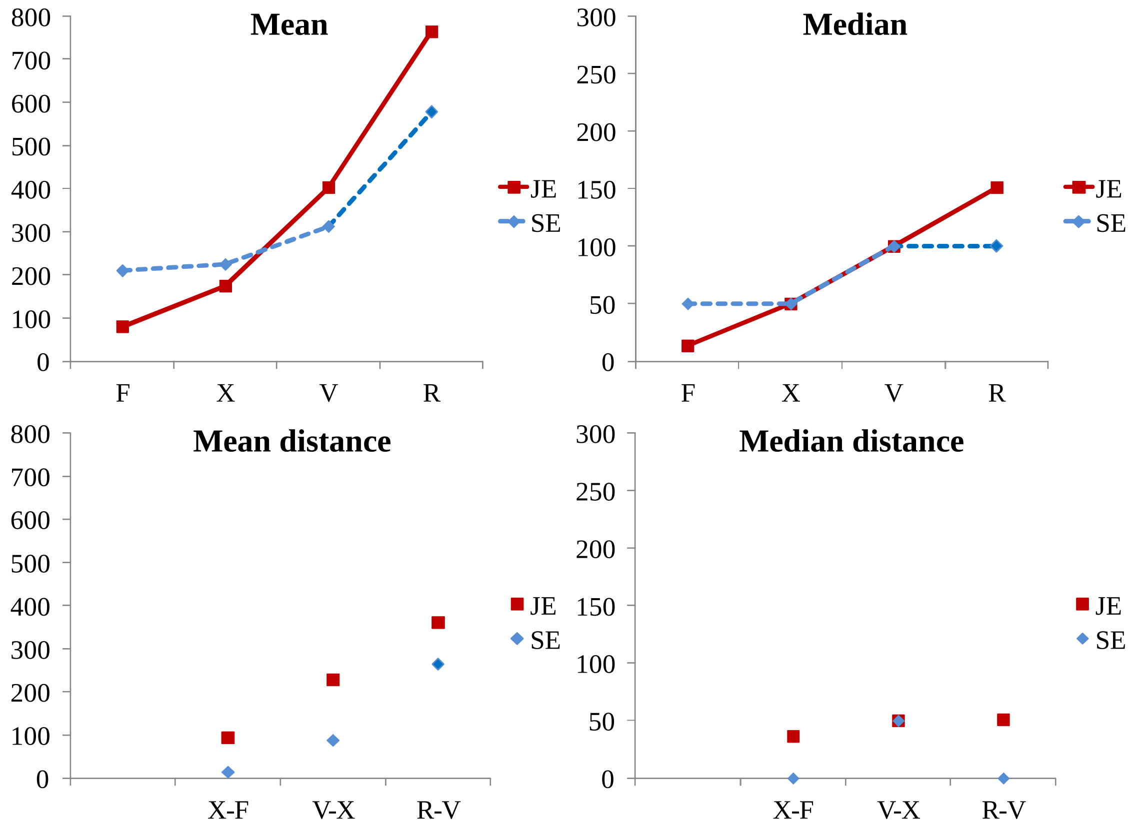


Figure 3. CV responses (€) in SE and JE and responses distance between adjacent NFRIs. Mean and median in JE are computed from all the groups that evaluate each NFRI[6]

---

we can compute utility of an injury as $I(w) = 1 - \frac{m_I}{m_D}$. Eventually, m_I and m_D are obtained from CV responses.

[6] Given the disproportionately high willingness to pay by some outliers, that unbelievably increase mean figures, those responses higher than the 99th percentile, for each group and NFRI, were dropped from the analysis. Performed Wilcoxon tests remain the same.

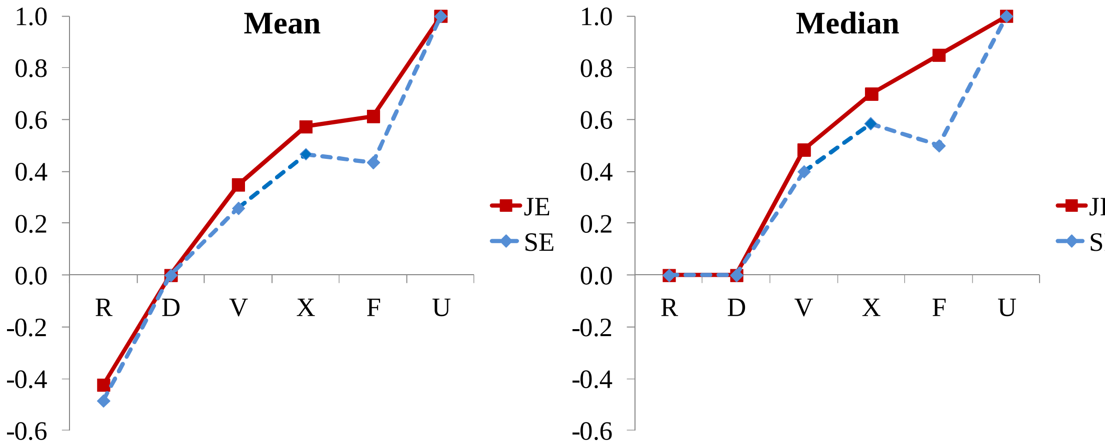


Figure 4. Utility function derived from CV responses in SE and JE

Now we are interested in analysing what characteristics of the respondents are related to a higher/lower value sensitivity in SE or JE. For that purpose in Table 2 we show the number of NFRIs pairs that has been evaluated statistically differently in SE and JE by different groups of respondents distinguished by some characteristics as gender, age, experience, driving habit, education, current monthly income, happiness, smoking and drinking, sports activity, and frequency of playing gambling games. Some of these deserve an explanatory comment. For example, experience is constructed from a question in which respondents report whether they, or some member of their family, have ever had a road injury. We build a happiness index as the average of responses to six questions on life satisfaction: Q1. *In most ways my life is close to my ideal*; Q2. *The conditions of my life are excellent*; Q3. *I am satisfied with my life*; Q4. *So far I have gotten the important things I want in life*; Q5. *If I could live my life over, I would change almost nothing*; Q6. *In general, I am happy*. Seven possible responses vary from 1, "*strongly disagree*", to 7, "*strongly agree*".

The number of NFRIs evaluated significantly different goes from 0, those groups with low value sensitivity, to 6, those groups with high sensitivity.[7] First, notice that in JE all the groups have evaluated differently the 6 pairs no matter their characteristics. So we can conclude that JE value sensitivity is not hindered by any factor. On the contrary, in SE some groups do better than others. People who are more value sensitive are those aged under 44, with experience on road injuries, who do not drive, with at least secondary education, above median income, above median happiness, who drink, do sports, and play gambling games only once a month or never. These results are consistent with GET which predicts that the three factors affecting evaluability (EM, experience, and nature) are *conjoint* in the sense that when one of these factors is in high evaluability level then the other factors do not make a difference on value sensitivity. Here, in JE (high evaluability) other factors do not affect sensitivity of responses but in SE they do. To further analyse this matter we perform econometric analysis to consider the effect of all this characteristics together in JE.[8] Our dependent variable is the difference in CV responses for

[7] Since we analyze 4 NFRIs we have the combinatory number of 6 pairs of HSs.
[8] Notice that this exercise is only possible in JE given the intrinsic characteristic of SE.

each NFRIs pair.[9] The higher this difference is, the higher the value sensitivity. We estimate median responses through quantile regression proposed by Koenker and Bassett (1978).[10] The pooled estimation (i.e. considering the difference in responses for all the NFRIs pairs together) is shown in second column of Table 3. All the variables considered seem to be statistically significant except for experience. Therefore although no characteristic affects qualitatively the value sensitivity in JE, as shown in Table 2, this is affected quantitatively, since the difference of CV responses between two NFRIs varies. We also perform the same estimation for each of the 20 pairs of NFRIs. In columns 3 and 4 of Table 3 it is shown the number of significant (positive and negative) coefficients found along these estimations. The characteristics considered have also a significant effect on the value sensitivity of some NFRIs pairs. For example, *driving* has a positive effect in the pooled model and its coefficient is also positive and significant in the estimation of 6 HSs pairs.

Table 2. Number of NFRIs pairs evaluated significantly different, at least at 10% or lower of error in CV. By EM and characteristics

Characteristics	SE	JE	Characteristics	SE	JE
GENDER			HAPPINESS		
Male	3	6	High	4	6
Female	3	6	Low	2	6
AGE			SMOKING		
<=44	4	6	Smokes	2	6
>44	2	6	Does not smoke	2	6
EXPERIENCE			DRINKING		
Road injury	4	6	Alcohol	4	6
No road injury	2	6	No alcohol	0	6
DRIVING			SPORT		
Drives	2	6	Yes	4	6
Does not drive	4	6	No	2	6
EDUCATION			LOTTERY/GAMBLING		
>= Secondary	5	6	Almost every week or more	0	6
<=Primary	1	6	Once a month or less	3	6
INCOME					
High	3	6			
Low	1	6			

[9] Notice that in JE we have 8 HSs and therefore the combinatory number of 28 HSs pairs could be possible. However, given the specific structure of our survey (Table 1) we only have 20 pairs evaluated in JE.

[10] We opt for this econometric technique given that mean estimations implausibly varies when considering outliers out of our analysis.

Table 3. Median estimation of sensitivity of responses in CV

Variables	Pooled Coef. (S. E.)	Estimation for each pair of NFRIs N. of significant coefficients +	 -
Gender (cons: Female)			
Male	-4.9*** (1.9)	0	4
Age			
Age	-1.3*** (0.3)	0	4
Age^2	0.01*** (0.003)	4	0
Experience (cons: no injury)			
Road injury	1.0 (1.8)	1	1
Driving (cons: does not drive)			
Drives	6.9*** (2.1)	6	0
Education (cons: <=primary)			
Secondary	8.3*** (2.3)	5	0
Vocational	17.0*** (3.8)	12	0
Tertiary	6.4** (2.8)	8	0
Income (cons: 0-1200€)			
1200-1800€	3.5 (2.2)	4	0
>1800€	8.2*** (2.5)	5	0
Happiness (cons: below med.)			
Above or equal to med.	6.2*** (1.8)	4	0
Smoking (cons: does not)			
Does smoke	8.0*** (2.0)	6	0
Drinking (cons: does not)			
Does drink	5.7*** (2.0)	6	0
Sport (cons: does not)			
Does practice	8.2*** (1.8)	5	0
Lottery/Gamb. (cons: does not)			
Does play gambling games	6.6*** (2.2)	7	1
Constant	22.1*** (7.7)	6	0
Observations	6,015		

Note 1. Estimation coefficient above Standard Error, the latter shown in brackets. Note 2. ***, **, and * mean coefficient is significant at 1%, 5% and 10% of error respectively.

4.3. Modified Standard Gamble

In Figure 5 it is presented mean and median MSG responses in both analysed contexts, SE and JE. Mean and median figures in JE are computed from all sample groups that evaluate HSs in that Evaluation Mode. We can also see mean and median distance between adjacent NFRIs as an approximation to the value sensitivity between this adjacent HSs. Given that HSs in the horizontal axis are ordered by severity it is expected a positive slope of the curves, that is subjects take higher risk in *Treatment B* for avoiding more severe injuries. With respect to the distance no a priori pattern is expected.

Visually, value sensitivity results similar in both evaluation modes since mean responses graphs are similar in both EMs and either do distances. The range between the mildest and worst NFRI (F and L) is similar in both cases. However, some differences about the value sensitivity along that range between JE and SE appear. For example, the mean distances between X-W and S-V look higher in SE, while the distances between N-R and L-N seem to be greater in JE. For median figures the differences are more pronounced: for example, median responses to S, R and N are considerably higher in SE.

The performance of Wilcoxon ranksum tests conclude that value sensitivity is encountered in both EMs (H1) since equality of distribution of responses to adjacent NFRIs is rejected (p-value<0.1). Only SE responses for F and W are not significantly different (p-value>0.1) which is consistent with previous results for CV and with literature evidence that attach to SE less value sensitivity.

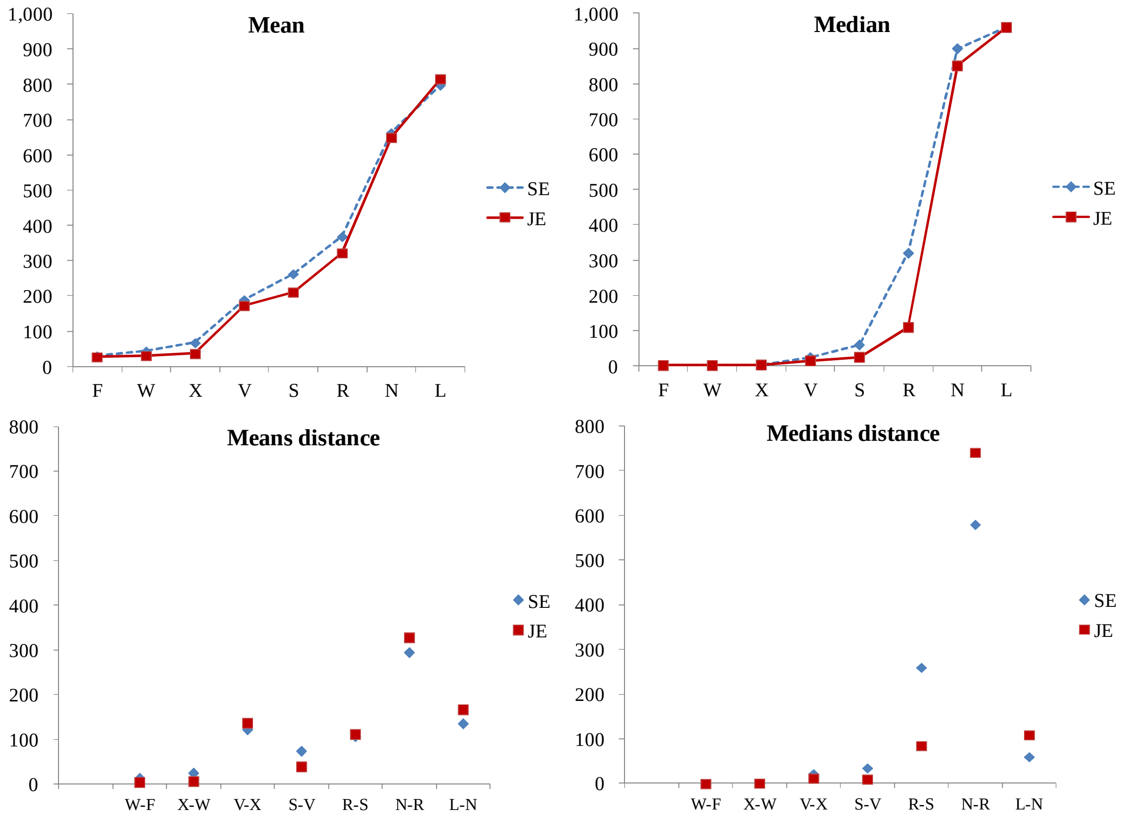


Figure 5. MSG responses (per thousand) in SE and JE and response distance between adjacent NFRIs. Mean and median in JE are computed from all the groups that evaluate each NFRI

Comparison of means and medians in Figure 5 is made between SE and JE of different subsamples. Given the structure of the survey (see Table 1) within group comparisons are available for MSG. This analysis is shown in Table 4. JE responses are significantly lower than SE answers for F, W and X, which are the mildest NFRIs. On the contrary, JE responses are significantly higher for L, which is the severest injury. These results suggest higher value sensitivity for JE mode as in previous evidence (see Hsee et al., 1999; Hsee et al., 2009; Bartels, 2006; or Chu, 2010). This leads to a higher range, distance between F and L for JE and also higher distance between X and V, and between N and L in JE. Notice that within group comparisons the same individuals that value the NFRIs in SE then significantly change their responses in JE. Another way of looking at the SE-JE change of responses is in the eighth column of Table 4 where it is shown the ratio of number of individuals with higher responses in SE to number of respondents with higher responses in JE. Higher responses in SE are relatively more (less) frequent for the milder (more severe) NFRIs which is consistent with Wilcoxon test results as expected.

Table 4. Joint vs Separate MSG responses (per thousand). Within-group comparisons

Group	**NFRI in SE**	$\pi_{separate}$		π_{joint}		**Wilcoxon Signrank test (z)**	$\frac{N_{\pi_{sep}>\pi_{joi}}}{N_{\pi_{sep}<\pi_{joi}}}$
		Mean	**Median**	**Mean**	**Median**		
1	F	30.3	2	19.9	2	2.68***	1.8
8	W	43.4	2	37.1	2	1.92*	1.5
3	X	67.6	3	39.6	3	2.30**	1.5
2	V	189.1	25	189.6	30	0.50	1.1
7	S	262.6	60	245.9	60	1.31	1.2
4	R	368.7	320	355.3	299	-0.14	0.9
6	N	662.9	900	647.2	851	-0.13	0.9
5	L	797.9	960	824.8	975	-3.01***	0.5

Note 1. Groups (first column) are ordered according to the severity of the injury evaluated in SE.
Note 2. ***, ** and * denote statistical significance at 1%, 5% and 10% respectively.
Note 3. $N_{\pi_{sep}>\pi_{joi}}$ ($N_{\pi_{sep}<\pi_{joi}}$) stands for number of individuals with higher responses in SE (JE) than in JE (SE).

In Table 5 we show the number of adjacent NFRIs pairs that has been evaluated statistically differently in SE and JE by different groups of respondents according to the same characteristics presented in the previous section.[11] The number of NFRIs evaluated significantly different goes from 0, those groups with low value sensitivity, to 7, those groups with high sensitivity.[12] MSG responses in JE have been more value sensitive for those respondents who do not smoke, play sports and do not use to play gambling games. More variables affect the value sensitivity in SE. In this EM it has been found higher evaluability among individuals under 44, those with experience, with high education, low income, low happiness, those who do not smoke, and

[11] We do not analyze the non adjacent NFRIs pairs because this has been evaluated statistically different for MSG in both EMs and independently of the characteristics of the individuals.
[12] Since we analyze eight NFRIs, that gives place to seven adjacent HSs pairs.

those with low frequency of playing gambling games. Again these results are consistent with GET since most of the factors do not make a difference on value sensitivity in JE. The econometric analysis for JE is shown in Table 6 where value sensitivity for the 20 HSs pairs is analysed.[13] A higher value sensitivity is found for those respondents with experience, higher education (only secondary) and income, and those who smoke. On the contrary those respondents who drive are associated with low sensitivity. It is noteworthy that here the experience arises as one factor with much importance even for the econometric analysis made separately for each HSs pair since in six of these estimations it is found to be significant with positive sign and in only one with negative sign. Finally, the significant effect of being male in the pooled estimation is not encountered in any of the 20 estimations for each pair of HSs.

Table 5. Number of adjacent NFRIs pairs evaluated significantly different, at 10%, 5% or 1% level, in MSG. By EM and characteristics

Characteristics	**SE**	**JE**	**Characteristics**	**SE**	**JE**
GENDER			HAPPINESS		
Male	5	7	High	5	7
Female	5	7	Low	6	7
AGE			SMOKING		
<=44	7	7	Smokes	4	6
>44	5	7	Does not smoke	6	7
EXPERIENCE			DRINKING		
Road injury	6	7	Alcohol	6	7
No road injury	5	7	No alcohol	6	7
DRIVING			SPORT		
Drives	6	7	Yes	6	7
Does not drive	6	7	No	6	6
EDUCATION			LOTTERY/GAMBLING		
>= Secondary	6	7	Almost every week or more	5	5
<=Primary	4	7	Once a month or less	7	7
INCOME					
High	5	7			
Low	6	7			

[13] Table 6 has the same interpretation than Table 3 in the case of CV.

Table 6. Median estimation of sensitivity of responses in MSG (per thousand)

Variables	Pooled Coef. (S. E.)	Estimation for each pair of NFRIs N. of significant coefficients +	 -
Gender (cons: Female)			
Male	1.2*** (0.4)	0	0
Age			
Age	-0.1 (0.1)	0	2
Age^2	6.9×10^{-4} (6.3×10^{-4})	2	0
Experience (cons: no injury)			
Road injury	1.2*** (0.4)	6	1
Driving (cons: does not drive)			
Drives	-2.1*** (0.4)	0	3
Education (cons: <=primary)			
Secondary	1.5*** (0.5)	1	0
Vocational	1.0 (0.8)	3	1
Tertiary	0.6 (0.6)	3	0
Income (cons: 0-1200€)			
1200-1800€	2.0*** (0.5)	3	0
>1800€	3.0*** (0.5)	3	0
Happiness (cons: below med.)			
Above or equal to med.	0.5 (0.4)	1	4
Smoking (cons: does not)			
Does smoke	1.1*** (0.4)	2	1
Drinking (cons: does not)			
Does drink	0.2 (0.4)	1	0
Sport (cons: does not)			
Does practice	-0.5 (0.4)	0	4
Lottery/Gamb. (cons: does not)			
Does play gambling games	-0.02 (0.5)	2	0
Constant	22.1*** (7.7)	14	0
Observations	6,015		

Note 1. Estimation coefficient above Standard Errors, the latter shown in brackets. Note 2. ***, **, and * mean coefficient is significant at 1%, 5% and 10% of error respectively.

4.4. Discussion

For both state preference elicitation methods the value sensitivity is higher in JE. This result is consistent to previous studies in different domains. However, value insensitivity encountered in SE is much more extreme in CV. While in SE-CV evaluation of F is not different to evaluation of X, and CV responses to V are not statistically different to responses to R, these NFRIs have been evaluated statistically different in SE-MSG. This implies that when CV is implemented in SE then most of the value sensitivity that appears in JE is missed. However for MSG even in SE high value sensitivity remains.

Previous studies (Jones-Lee et al., 1995) for the valuation of NFRIs with Contingent Valuation and Standard Gamble shows that the former method was subjected to a high insensitivity consisting on individuals showing the same willingness to pay for different size of risk reduction and different HSs with varying seriousness. The results presented here show that this insensitivity could be even higher in a SE mode but mostly disappear in JE. If these authors considered the value sensitivity of MSG as the most prominent feature for supporting it as the best evaluation procedure for valuation of NFRI, now these results provide a new good characteristic of this methods given its consistency to change in the evaluation mode, either in SE or JE MSG responses are sensitive to severity of the HSs to evaluate.

The implication of using SE rather than JE for the evaluation of NFRIs is the same as shown in previous works in other domains: best outcomes have a higher value in JE and worst outcomes have a higher value in SE. For example, in CV when F is evaluated in JE respondents just pay, on average, €80 for its prevention, however when F is evaluated in SE then average willingness to pay is €210. The same happens in MSG, when the evaluation mode is JE respondents are less willing to risk their lives to avoid F injury than when the evaluation is separate. The implications for policy decision making is that if we consider the JE as the proper evaluation mode then values that may have been elicited in a SE framework should be updated or corrected in order to eliminate insensitivity bias. This correction should be upward for the value of preventing the more serious injuries and downward for the value of mild injuries. Of course, this correction should be different in its intensity depending on the method of evaluation considered, MSG or CV.

5. Testing consistency of responses across contexts in Joint Evaluation

5.1. Range-frequency effects

According to Range-Frequency Theory (RFT, Parducci, 1965) the internal judgment of an injury is governed by two principles: the range and the frequency principle. Specifically the range principle is concerned with the proportion of the psychological range, the difference between the two extreme values included in a context, that is below the value of an injury. The frequency process is such that the value attached to an injury is given by the proportion of contextual stimuli below the target stimulus (the injury being valued) in the specific context.

Formally the range value of avoiding a risk of an injury *i* in an context *j*, $Range_{ij}$, is given by (see Wedell and Parducci, 1988):

$$Range_{ij} = \frac{V_i - V_{min,j}}{V_{max,j} - V_{min,j}} \tag{2}$$

Where $V_{max,j}$ and $V_{min,j}$ are the maximum and minimum values included in context j and V_i is the "context-free" (see Robinson et al., 2001) value of avoiding the injury.

For the frequency value, $Freq_{ij}$, is given by the next expression (see Wedell and Parducci, 1988):

$$Freq_{ij} = \frac{rank_{ij} - 1}{N_j - 1} \tag{3}$$

Where $rank_{ij}$ is the rank of the injury within its context. For example, the rank of F in group one and five is 1 and the rank of L is 4 (see Table 1). N_j is the number of NFRIs in the context. In our study $N_j = 4$ for $j=1,\ldots,8$.

Finally, the integration of these two evaluation processes constructs an internal judgment. The final subjective value of avoiding the risk of the injury is given by a weighted average of the two values:

$$J_{ij} = \omega \times Range_{ij} + (1 - \omega) \times Freq_{ij} \tag{4}$$

With ω being the relative importance of the two principles. When $\omega = 1$ then the range principle is the only process that affects valuation depending on the context. However, if $\omega = 0$ then the subjective value is totally given by the rank of the NFRI in a specific context. Notice that since we are interpreting J_{ij} as the subjective value of avoiding a risk of injury this should be higher for severe injuries and consequently the willingness to pay or to risk for avoiding them should also be higher.

In our survey we have the same injury evaluated in different contexts that differ in the extreme contextual values included. However each NFRI has the same rank across the groups where presented. This way we can only check for range effects by comparing responses to the same injury in different contexts. In Table 7 we can check range values for each NFRI in different groups. There are some injuries that are expected to have a different subjective value because their range value changes over contexts. These are shown in bold: W, V, S and N. For example, the range value for W is lower in groups 1&5 than in 4&8 because in the former case the maximum value (the severest injury) is higher; injury L rather than injury R. On the contrary, according to expression (2) the range value of F, X, R and L remains the same among the groups. Notice that this is because the latter injuries are those that are extreme values in each context. For example: F (X) is the maximum value in groups 1, 4, 5, and 8 (2, 3, 6 and 7); R (L) is the maximum value in groups 3, 4, 7, and 8 (1, 2, 5 and 6). Notice also that the range values are the same in JE for MSG and CV since the injuries included in the evaluation are the same for each group (see Table 1). So it is expected the same range effects in both elicitation methods.

We can also consider frequency effects by comparing different NFRIs like X and W. In JE the final subjective value of avoiding X (expression 4) is zero given that it is the best health state in any of the contexts where it is evaluated (in groups 2, 3, 6 and 7 its rank is 1). However W, a less severe health state, is not the best health state among the NFRIs that are evaluated jointly with it (in groups 1, 4, 5 and 8 its rank is 2) and has a range and frequency value higher than zero. This leads to $J_{W,(1,4,5\&8)} > J_{X,(2,3,6\&7)}$ (or $J_{W,(5\&8)} > J_{X,(6\&7)}$ in the case of CV). The prediction in this case is that we could find higher MSG and CV responses for W even though it is a less severe injury than X. The same argument could apply to R and N. The former has a range and frequency value of 1 (in groups 3, 4, 7 and 8 its rank is 4) while the latter is lower than the unity (in groups 1, 2, 5 and 6 its rank is 3). The final subjective value of avoiding N is lower than the value of avoiding R even though it is an objectively more severe injury. This is $J_{R,(3,4,7\&8)} > J_{N,(1,2,5\&6)}$ (or $J_{R,(7\&8)} > J_{N,(5\&6)}$ in the case of CV). Notice that we do not strictly expect higher willingness to pay (or risk) for W and R than for X and N, respectively. This only happens in the extreme case that a respondent follows exactly RFT. At the end of the day, if we find that responses for X and N are higher than W and R respectively this is not evidence for rejecting some effects driven by range and frequency value (of course, it would be evidence against individuals strictly following RFT). However, if we find that W and R responses are higher than X and N ones, respectively, then it is evidence for the existence of some RFT effects because in the absence of this effects the expected responses are the other way around.

Table 7. Range values in JE for MSG and CV

NFRIs	Groups						
	1 & 5		2 & 6		3 & 7		4 & 8
F	$Range_{F,1\&5}$			=			$Range_{F,4\&8}$
W	$\boldsymbol{Range_{W,1\&5}}$			<			$\boldsymbol{Range_{W,4\&8}}$
X			$Range_{X,2\&6}$	=	$Range_{X,3\&7}$		
V			$\boldsymbol{Range_{V,2\&6}}$	<	$\boldsymbol{Range_{V,3\&7}}$		
S					$\boldsymbol{Range_{S,3\&7}}$	<	$\boldsymbol{Range_{S,4\&8}}$
R					$Range_{R,3\&7}$	=	$Range_{R,4\&8}$
N	$\boldsymbol{Range_{N,1\&5}}$	>	$\boldsymbol{Range_{N,2\&6}}$				
L	$Range_{L,1\&5}$	=	$Range_{L,2\&6}$				

Note. For CV we only consider groups from 5 to 8.

5.2. Anchoring effects

In the literature of health utilities elicitation Stalmeier (2002) showed that the existence of anchor effects could be a possible reason for the inconsistencies encountered in utilities elicited by Time Trade-off (TTO) when the reference health state is changed. The explanation relies on the fact that responses to TTO questions are positive related to the anchor. In our present framework more severe injuries should obtain higher responses (i.e. higher WTP or WTR) then when the evaluation of an injury is preceded by a severe injury the anchor is higher than when preceding by a mild injury and then higher responses are expected in the former case. The anchoring effects that we expect for MSG and CV in different context are shown in Table 8. For example, injury F in groups 1 & 5 can be preceded by W, N or L, while in groups 4 & 8 can be preceded by W, S or R. Given that the order of valuation of each injury is random, it can be computed that there is about 38% of probability that F is preceded by N or L in groups 1 & 5 while F is preceded by S or R in group 4 & 8. Since S and R are less severe than N and L we expect (according to anchoring effects) that responses to MSG and CV are higher in groups 1 & 5. Notice that some of the anchoring effects are opposite to the Range Effects considered in Table 7. In those cases, it is interesting to see which effect predominates.

Table 8. Effect of anchor on responses (R_{ij}) in JE for MSG and CV

NFRIs	Groups 1 & 5		2 & 6		3 & 7		4 & 8
F	$R_{F,1\&5}$			>			$R_{F,4\&8}$
W	$R_{W,1\&5}$			>			$R_{W,4\&8}$
X			$R_{X,2\&6}$	>	$R_{X,3\&7}$		
V			$R_{V,2\&6}$	>	$R_{V,3\&7}$		
S					$R_{S,3\&7}$	>	$R_{S,4\&8}$
R					$R_{R,3\&7}$	>	$R_{R,4\&8}$
N	$R_{N,1\&5}$	<	$R_{N,2\&6}$				
L	$R_{L,1\&5}$	<	$R_{L,2\&6}$				

Note. For CV we only consider groups from 5 to 8.

5.3. Results[14]

Mean and median responses for CV and MSG in each context are presented in Table 9 and 10 respectively. First, we comment on CV responses. We find context effects for the valuation of five HSs. WTP for avoiding a risk of F is higher in group 8 than in group 5 while no Range-effect is expected and the anchoring effect is precisely in the opposite direction. Responses to X and V are significantly higher in group 7 than in group 6, which is inconsistent with the anchoring effect expected, although in the case of V the results are consistent with the Range-effect (see Table 7). Eventually, injuries S and R have been valued as worse injuries in group 7 than in group 8 which is consistent with the anchoring effects predicted but, in the case of S, inconsistent with range effects (see Table 7).

In MSG only the evaluation of X and S presents contexts effects. In the case of X mean responses are higher in groups 3 & 7 than in groups 2 & 6 which is inconsistent with the anchoring effect expected (the expected effect is opposite, see Table 8). With respect to S higher responses in groups 3 & 7 than in 4 & 8 arise which is consistent with anchoring effect but not with Range-effect shown in Table 7. Notice that the anchoring effect is the dominant effect in the evaluation of injury S in MSG and CV. Although in the case of Contingent Valuation of injury V the Range-effect predominates.

As a general conclusion we have that contexts effects is largely presented in CV but not in MSG. Nevertheless, found contexts effects do not match theoretical predictions in Table 7 and 8. In case of conflicting predictions (in the evaluation of W, V, S and N) between the two principles here analysed (anchor and Range Value) results can be explained by the predominant effect. However in case of F, X, R and L no conflicting prediction exists and despite of that we do not find the expected effect (except for R in CV).

Eventually, we find evidence of theoretical predictions about frequency effects for CV. WTP for avoiding a risk of R in groups 7 and 8 (mean=€1,056.2, median=€160, p25=€50, and p75=€601) is statistically higher than WTP for avoiding a risk of N in groups 5 and 6 (mean=€822.6, median=€150, p25=€30, and p75=€600) with a Ranksum-test p-value=0.04 denoting that WTP responses for avoiding R are shift above WTP for avoiding N. Since N is a more severe injury WTP should be higher. However we think that these results are given by a frequency effect because the rank of R is higher than the rank of N in their corresponding contexts. In other words, we find evidence for $J_{R,(7\&8)} > J_{N,(5\&6)}$ (see section 5.1 above). On the other hand the responses to the evaluation of W in groups 5 and 8 (mean €165 and median €50) are significantly lower to the responses to the evaluation of X in groups 6 and 7 (mean €174.5 and median €50) with a Ranksum-test p-value=0.05. So no evidence of frequency effect is found in the evaluation of W and X. We do not find frequency effects in the case of MSG. WTR for avoiding W in groups 1,4,5&8 (mean=42.5, median=2, p25=2, and p75=4) is statistically lower than WTR for avoiding X in groups 2,3,6&7 (mean=41.2, median=3, p25=2, and p75=5) with a Ranksum-test p-value=0.008.[15] WTR for avoiding R in groups 3,4,7&8 (mean=312.7,

[14] We drop out the 99th percentile, for each group and NFRI, as in JE-SE analysis above (see footnote 6). However when we report mean or median based on two groups (e. g. mean CV responses for groups 7 and 8) we drop out the 99th percentile with respect to these two groups. In any case performed Wilcoxon tests do not change.

[15] Although mean WTR for avoiding W is higher than WTR for avoiding X, the analysis of percentiles and the results of Wilcoxon ranksum test indicate the opposite. For example Wilcoxon rank sum (necessary to compute the W-ranksum statistic) for W (X) responses is lower (higher) than expected.

median=101, p25=4, and p75=551) is statistically lower than WTR for avoiding N in groups 1,2,5&6 (mean=664.9, median=901, p25=451, and p75=980) with a Ranksum-test p-value<0.001.

Table 9. Context effects in JE. CV

NFRIs	**Groups**				**Wilcoxon ranksum test**
	5	**6**	**7**	**8**	**(*z*)**
F	71.3 11			89.6 30	-2.12**
W	112.2 40			178.7 50	-1.49
X		139.0 30.0	210.2 60.0		-3.92***
V		348.7 65.0	457.4 100.0		-3.65***
S			602.4 150.0	514.7 100.0	2.68***
R			1063.7 200.0	741.4 147.5	2.82***
N	792.2 150.0	1055.7 150.0			-1.17
L	821.4 177.5	1439.0 160.0			-0.6

Note. Mean (above) and Median (below) in €.

Table 10. Context effects in JE. MSG

NFRIs	Groups 1 & 5	2 & 6	3 & 7	4 & 8	Wilcoxon ranksum test (*z*)
F	31.4 2.0			23.1 2.0	0.10
W	34.2 2.0			27.7 2.0	-1.23
X		26.4 3.0	46.7 3.0		-2.31**
V		174.1 15.0	171.2 15.0		-0.11
S			232.3 50.0	189.8 12.0	3.42***
R			324.0 144.5	320.0 100.0	0.91
N	659.6 900.0	639.0 820.0			0.42
L	814.5 955.0	816.7 960.0			0.82

Note. Mean (above) median (below) in per-thousand units.

5.4. Discussion

Valuation of safety seems to be affected by context effects far more in the case of CV than in the case of MSG. This result give place to several conclusions that complement the analysis performed in section 4. In the first place, it is shown again that MSG is a more invariant method to several manipulations of the elicitation process. In section 4 it is shown that it is not affected by the evaluation mode; at least not to the extent that CV is affected. What is more in this section we find that MSG is mainly invariant to context manipulation. On the other hand, CV is affected by both Evaluation Mode and contextual injuries. Therefore the expectation that CV can be improved when provided in a Joint Evaluation fashion, because its value sensitivity is enhanced in that case, is hindered by the fact that it is in this evaluation mode where the valuation of an injury depends on the remaining health conditions that are valued jointly. So JE seems to be a better evaluation mode, however it is subjected to context effects.

Moreover, when a systematic bias occurs it is possible to compute that bias so that elicited preferences can be bias-corrected. However, the context effects found do not seem to follow a clear pattern. They are not totally consistent with range effects predicted by RFT. Neither anchoring effects is a good explanation. Under these circumstances we cannot apply bias correction to obtain individual preferences. Eventually we have several valuations depending on the contextual injuries so that a decision has to be made to choose which valuation is the more appropriate. In this sense we need to find which context is suitable to elicit preferences. Normative criteria for valuation could help in this task. For example, one criterion can be to consider that the valuation context should be as close as possible to the real context of road safety. This requires that the context is formed by those injuries that are frequent in real situations.

6. Valuation of road safety under reference dependent evaluation of income

In the survey questionnaire two different questions are included to measure two income concepts, namely current income and permanent income. The former refers to the amount of money earned currently, while the latter has to do with a long term concept of income like the average income throughout a whole life.

First respondents were asked about their current income as follows:

…Regarding the level of income of your household, and approximately, could you mark the interval that correspond to your situation?

- *a) Less than €600.*
- *b) Between €601 and €900.*
- *c) Between €901 and €1,200.*
- *d) Between €1,201 and €1,800.*
- *e) Between €1,800 and €2,500.*
- *f) Between €2,501 and €3,500.*
- *g) Between €3,501 and €5,000.*
- *h) More than €5,000.*

Then they were asked about their permanent income as:

…As you may know, over the life of an individual different stages in terms of income occur (sometimes you earn a lot, others less). When we consider these various stages, people are able to identify a "normal income level" throughout our entire lives. This results in that you may think that your current income level is above or below its "normal income level." We would like you to tell us what would be, among the following, your "normal income level"?

- *a) Less than €600.*
- *b) …*

This is a question that require respondents to think about their past income and future income to elaborate their “normal income throughout their entire lives”. Our claim is that respondent reports their permanent income or “normal income” as a reference point to evaluate their current income rather than as a measure of their financial capacity.[16] Given this two income concepts we can identified those who are in a loss frame, current income is lower than normal income, those who are in a gain frame, current income is higher than normal income, and those who are in a neutral position, with current income equal to normal income. The majority, 1,285 out of 2,016 respondents, report their current income to be the same as their normal income. The second most frequent frame is loss, 633 individuals have current income below their permanent income. Eventually, 98 respondents are in a frame of gains. We believe that this asymmetry in frequency of frames is due to the economic crisis presented in Spain since 2008 (three years before the survey was carried out).

6.1. Reference dependent evaluation of income

In this section we analyse if the framing just defined has an effect on life satisfaction. To see this we compute the mean of the Happiness Index describe in Section 4.2 for each frame and each current income level. In Figure 6 it is shown that the framing has an effect on happiness in such a way that those respondent in the gain frame (normal income<current income) report a higher mean of Happiness Index than those in a neutral position (normal income=current income) and the latter are also happier than those in a loss frame (normal income>current income).[17] We also test the null hypothesis of equality of distribution of the Happiness Index by performing Wilcoxon-Raksum tests. For example, among those with current income between €1,201 and €1,800, those in a gain frame have a distribution of happiness index significantly (at 10%) different to those in a neutral frame (z statistic=-1.93), the latter have a different distribution (sig. at 5%) to those in a loss frame (z statistic=-2.39), and finally those in a gain scenario also have a different distribution of the Happiness Index (sig. at 5%) to respondents in a loss frame (z statistic=-2.49). We find that the distribution of Happiness Index differ significantly between frames for the other intervals of current income as well, except for the interval €1,801-€2,500. In Figure 7 we can see that this pattern is also encountered when we take into account Self-Reported Happiness in different type of life satisfaction questions. In most of the income groups those in gain frame report higher happiness than respondents in a neutral frame and these latter are also happier than those in a loss frame. Only responses to Happiness question 5 seem to not follow this pattern to some extent.

To further analyse the effect of the framing on happiness we perform an econometric analysis that allows us to compute an effect of framing having into account all the observations that are not presented in Figures 6 and 7 (see footnote 17). Also the econometric analysis allows us to control for several variables that may affect Self-Reported Happiness, namely: age, gender, self-reported health, marital status, number of minor children at home, dependent elderly at home, employment status and education. Given the categorical nature of responses to the happiness questions we decided to estimate a Probit model, specifically we estimate the probability that a response is equal or above the median response. In Table 11, the estimated coefficients are shown when the dependent variable is the Happiness Index (second column) and different type of Self-Reported Happiness (columns 3 to 8). For the Happiness Index model, it is increasing with current income since we find the current income dummies to be significant (except for

[16] We use permanent income, the theoretical concept, and normal income, the word presented to respondents, as synonyms.

[17] Notice that this figure do not show individuals in other income intervals as those with less than €900 or those with more than €3,500. This is because number of respondents in some frames is less than 5 and we believe that no proper analysis can be done in such cases.

intervals 901-1,200 and 2,501-3,500). Notice that the effect of current income is conditional to the framing being constant so we can interpret as the effect of incrementing current income jointly with normal income (so that the frame does not change). Going to the effect of framing we estimate those in neutral and gain frame to be happier than those in loss frame (positive and significant coefficients). Indeed, those in the gain frame are happier than respondents in neutral frame (see equality of coefficients test in *Gain – Neutral* dummy). Estimation of different happiness questions shows mainly the same conclusion about the effect of framing; although not always significant the effect goes in the same direction. Also in these models happiness seems to be positive related to current income except for happiness question 3 and 5.

Permanent income hypothesis (Friedman 1957 and 1963) suggests that individuals should take into account their permanent income in their consumption decisions and therefore this should reflect a varying happiness pattern. The evidence that we find is somehow surprising in the sense that conditional on current income those who have a lower (higher) permanent income are supposed to have less (more) financial capacity to maintain their well-being which at the end of the day should be the determinant of happiness. However happiness is higher (lower) for this group. The results are consistent with a theoretical model in which both absolute income and aspiration levels (reference income) affect SWB as proposed by Easterlin (2001). In this model, the income satisfaction for a given level of income is lower (higher) if the reference point is high (low). In our case the reference point would be given by the permanent income. So the results here presented imply that permanent income is negatively related to life satisfaction. This result do not contradict permanent income hypothesis per se because it can be explained by the fact that happiness is not related to well-being (given by consumption) but to changes in well-being or specifically to the location of the income with respect to a reference point. Nonetheless in the next section we test if the framing has an impact on CV and MSG responses.

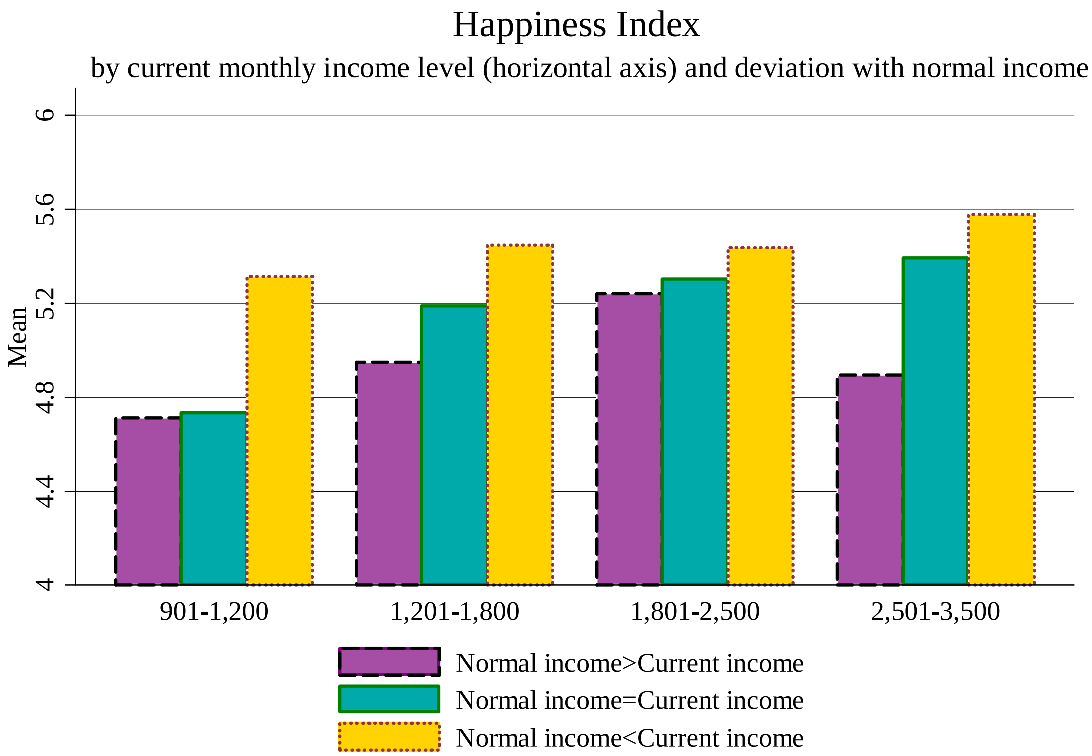


Figure 6. Mean of Happiness index by different frames

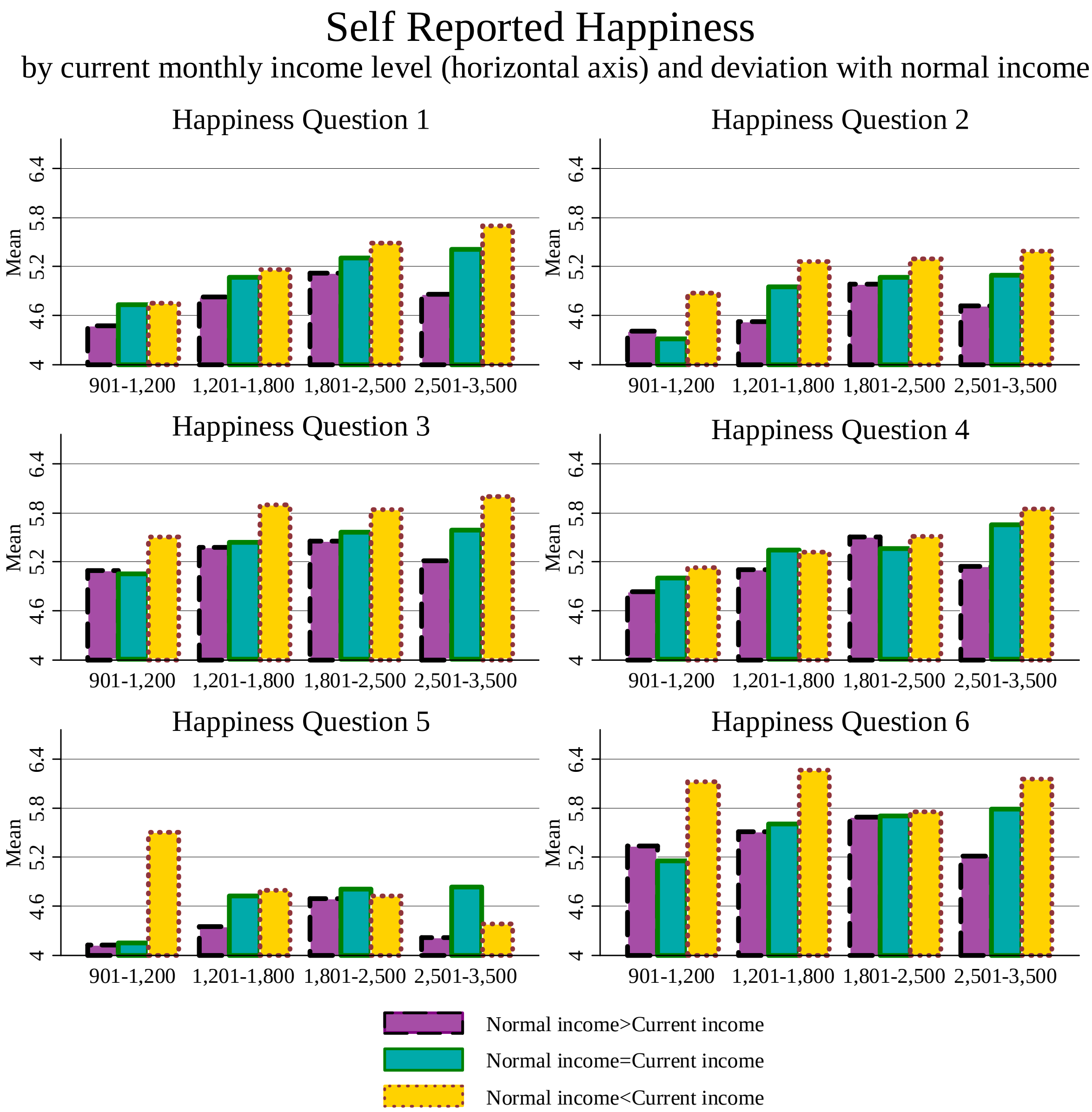


Figure 7. Responses to life satisfaction questions by different frames

Table 11. Estimation of Life satisfaction. Probit model

Variables	Happ. Index	Type of Self-Reported Happiness					
		Hap. Q.1	**Hap. Q.2**	**Hap. Q.3**	**Hap. Q.4**	**Hap. Q.5**	**Hap. Q.6**
Current Income (€, cons: <=900)							
901-1,200	-0.04	-0.06	-0.02	-0.03	-0.03	-0.22***	-0.10
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
1,201-1,800	0.21**	0.11	0.42***	0.15	0.14	0.13	0.18**
	(0.09)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
1,801-2,500	0.27**	0.28**	0.45***	0.16	0.22**	0.22**	0.19*
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
2,501-3,500	0.22	0.40***	0.40**	0.17	0.37**	0.18	0.22
	(0.15)	(0.15)	(0.16)	(0.15)	(0.15)	(0.15)	(0.15)
>=3,501	0.42**	0.60***	0.56**	0.34	0.36*	0.23	0.32
	(0.21)	(0.21)	(0.22)	(0.21)	(0.21)	(0.20)	(0.21)
Framing (Cons: Loss, NI<CI)							
Neutral (NI=CI)	0.19***	0.19***	0.13**	0.10	0.14**	0.15**	0.06
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Gain (NI<CI)	0.51***	0.27*	0.22	0.49***	0.45***	0.09	0.44***
	(0.15)	(0.15)	(0.16)	(0.16)	(0.15)	(0.15)	(0.16)
Gain - Neutral	0.32**	0.08	0.09	0.39***	0.31**	-0.05	0.38**
	(0.15)	(0.14)	(0.15)	(0.15)	(0.14)	(0.14)	(0.15)
Constant	0.89**	0.70*	0.72*	1.06***	0.47	1.94***	1.85***
	(0.38)	(0.39)	(0.37)	(0.37)	(0.37)	(0.37)	(0.37)
Observations	2,016	2,016	2,016	2,016	2,016	2,016	2,016

Note 1. Gain – Neutral are the coefficient for those in gain frame when the constant is those respondents in neutral frame. Note 2. These estimations are also controlling for age, gender, self-reported health, marital status, minor children, dependent elderly at home, employment status and education. Note 3. Estimated coefficients above Standard Errors, the latter shown in brackets. Note 4. ***, **, and * mean coefficient is significant at 1%, 5% and 10% of error respectively.

6.2. Effect of reference income on CV responses

In Figure 8 we can see median WTP for the prevention of a risk of road fatality by current income and framing.[18] Interestingly the pattern that we find is that those individuals in a gain frame are willing to pay more than those in a neutral fame. Also those individuals in a loss frame have higher median CV responses than those whose current income and normal income coincides. In the case of the interval of income €1,501-€2,500, Wilcoxon-Ranksum test rejects the equality of distribution of CV responses of individuals in frame of gains with respect to: those in neutral (z statistic=-2.45), and; those in a loss frame (z statistic=-1.85). Also for the case of those with current income between €2,501 and €3,500, we can reject at 1% the hypothesis of equality of distribution of responses between respondents in a frame of losses and those in the neutral situation (z statistic=2.63). In Figures 9 and 10 we show WTP for the prevention of a risk of different injuries again distinguishing by current income and deviation with normal income (framing).[19] In those figures we can find for injuries X, V, S, R, N, and L, the same pattern, i.e. in general respondent in gain and loss frame report a higher WTP than those in the neutral situation. However we cannot see a systematic pattern in the case of F and W.

In Table 12 we show the results of the median estimation of CV responses for preventing a risk of death and each injury separately. This estimation is controlling for the same variables than in Table 11. With this specification we can estimate the effect of current income (conditional of framing being constant) and the effect of the reference point of income (i.e. the effect of the framing) on the willingness to pay for road safety. With respect to the WTP to prevent a risk of fatal accident current income has a positive and highly significant effect (all dummies are significant at 1%). Also we find the same pattern presented in Figure 8 with respect to the effect of framing. The negative and significant coefficient for those individuals with normal income equal to current income (neutral frame) means a lower WTP than individuals in the constant (those in a loss scenario). We also find higher WTP in the gain frame with respect to both: the loss frame (positive coefficient for the gain frame dummy) and neutral frame (positive and significantly different coefficients of the dummies for the gain and neutral frame). Eventually, estimation of WTP for preventing a risk of each injury shows broadly the same result for the effect of current income; a positive effect. However, the effect of framing changes somehow. The most frequent coefficient for the frame of gains is positive (five out of eight injuries). Therefore we find the same effect as for WTP for preventing a fatality. Also for five injuries we find that those in the gain frame have significantly higher median CV responses than respondents in the neutral frame. Meanwhile the most common coefficient of the neutral frame is not significantly different from zero, which suggests no differences with respect to the loss frame.

[18] We opt for showing the median CV responses because mean figures are extremely sensitive to the inclusion or removal of some outliers.

[19] Given that the evaluation of the injuries was carried out by smaller subsamples (see Table 1), in Figure 9 and 10 we also consider all those intervals of income in which there is at least 5 observations for at least two frames.

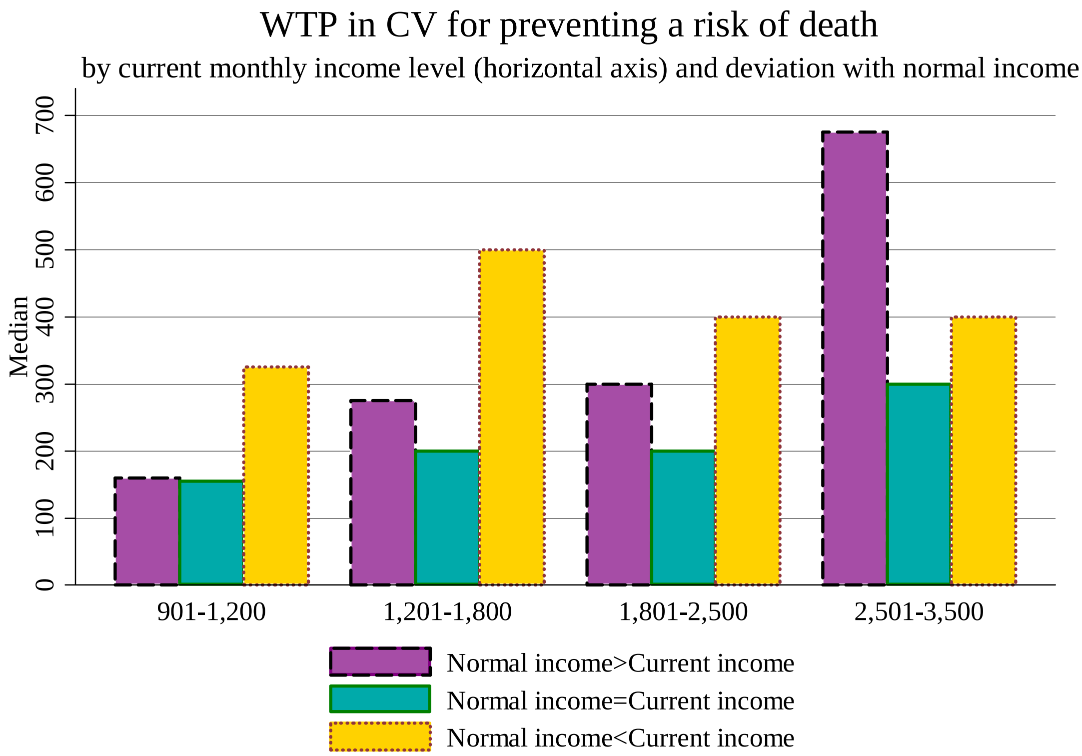


Figure 8. CV responses (€) by different frames. Prevention of death

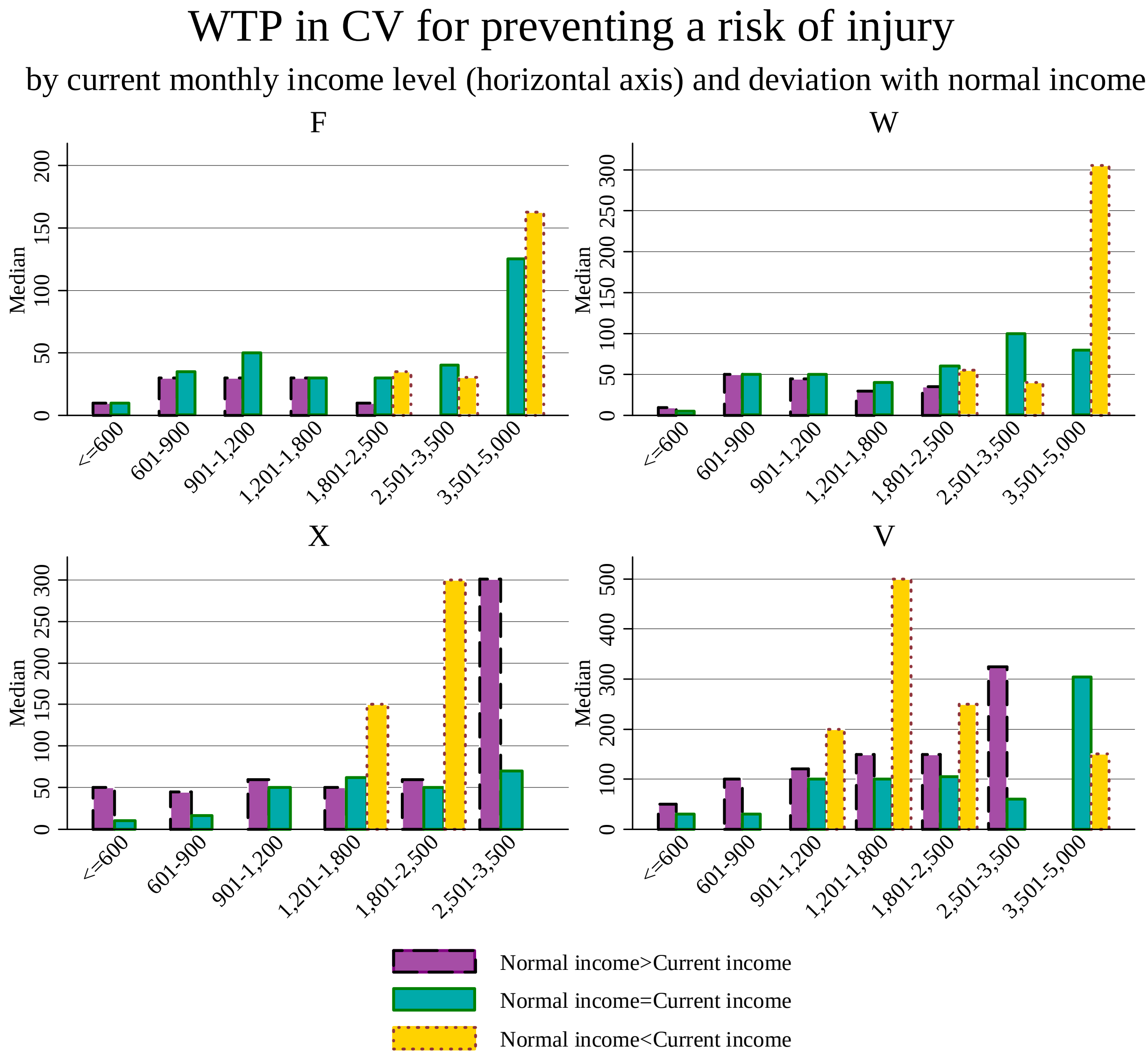


Figure 9. CV responses (€) by different frames. Prevention of injuries. (Part A)

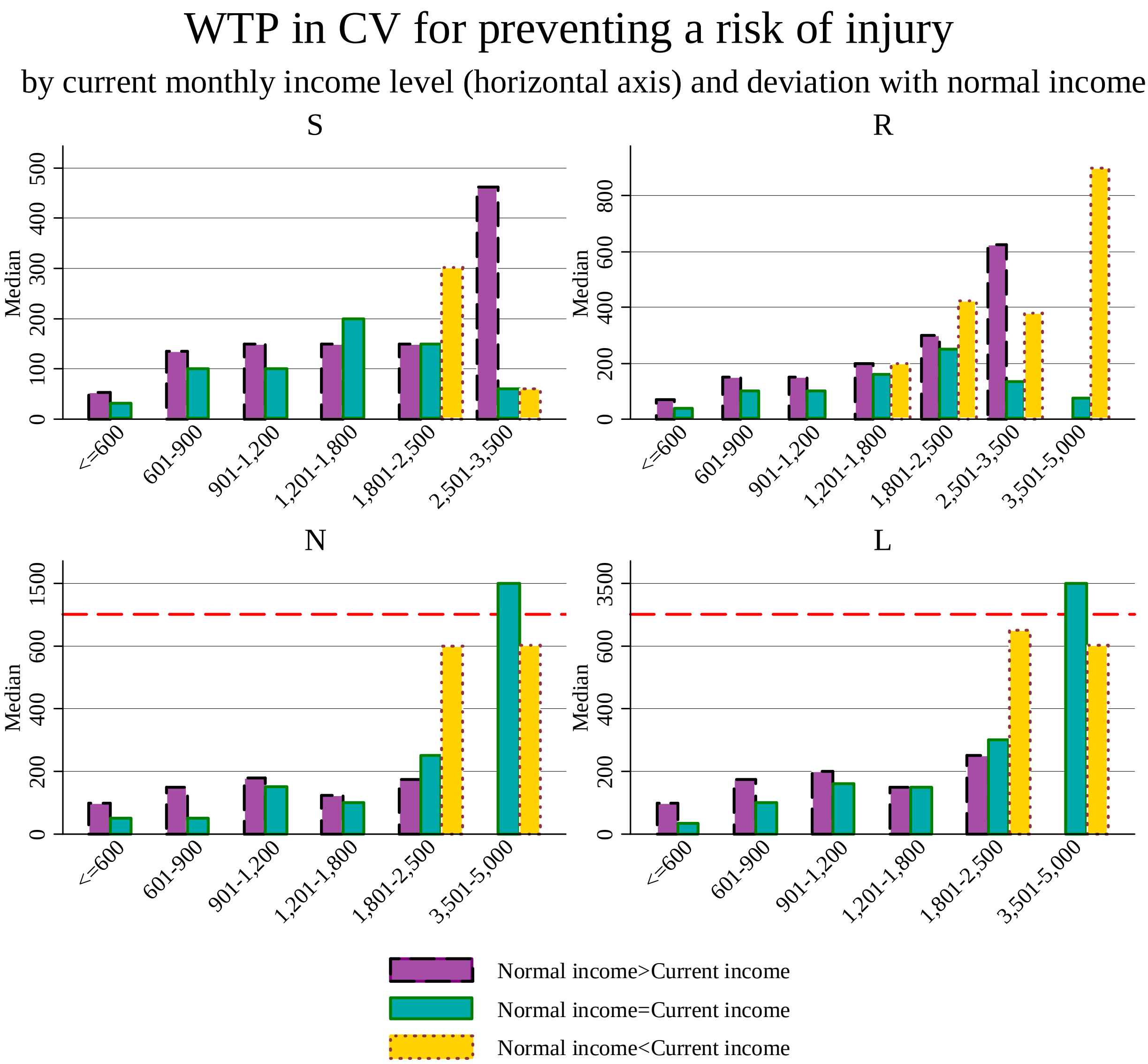


Figure 10. CV responses (€) by different frames. Prevention of injuries. (Part B)

Table 12. Median estimation of CV responses (€)

Variables	CV response (WTP for preventing)								
	Death	F	W	X	V	S	R	N	L
Current Income (€, cons: <=900)									
901-1,200	66.1***	12.6**	7.3	30.6**	44.7**	23.6	5.5	56.4	55.9
	(16.4)	(5.7)	(7.0)	(14.6)	(20.3)	(16.7)	(34.7)	(46.2)	(44.1)
1,201-1,800	90.8***	-4.4	-7.5	39.4**	47.4**	86.3***	52.6	5.4	13.4
	(17.9)	(6.2)	(7.4)	(15.8)	(22.6)	(18.2)	(38.0)	(49.7)	(47.9)
1,801-2,500	93.9***	-7.5	1.9	32.5*	57.1**	44.2**	98.0**	53.8	149.9**
	(21.0)	(7.3)	(9.2)	(18.9)	(25.8)	(21.4)	(44.1)	(64.3)	(63.2)
2,501-3,500	202.6***	2.3	7.1	68.5**	45.7	-5.0	27.3	158.2*	-13.6
	(28.9)	(10.2)	(13.2)	(27.2)	(36.1)	(30.0)	(58.0)	(91.6)	(92.6)
>=3,501	332.3***	108.1***	221.9***	43.4	125.8**	452.8***	377.1***	601.7***	682.85***
	(38.9)	(13.5)	(15.6)	(34.1)	(53.2)	(41.7)	(84.3)	(107.7)	(103.7)
Framing (Cons: Loss, NI<CI)									
Neutral (NI=CI)	-24.6*	4.5	21.0***	-14.5	-14.7	-28.7**	-40.0	29.3	34.7
	(12.6)	(4.3)	(5.1)	(11.4)	(15.9)	(12.2)	(25.5)	(37.2)	(35.9)
Gain (NI<CI)	194.0***	12.9	4.1	101.2***	115.7***	78.4***	146.9***	254.3***	96.3
	(28.5)	(10.6)	(12.2)	(29.9)	(33.0)	(26.0)	(54.1)	(96.8)	(91.3)
Gain - Neutral	218.6***	8.4	-16.8	115.8***	130.5***	107.1***	187.0***	257.3***	61.6
	(26.9)	(10.0)	(11.5)	(28.4)	(30.7)	(24.9)	(51.0)	(86.0)	(87.6)
Constant	99.6	44.7*	-17.7	48.8	46.5	467.8***	255.3*	-8.0	-89.8
	(71.4)	(24.7)	(29.8)	(65.5)	(91.0)	(73.1)	(150.2)	(212.4)	(202.7)
Observations	2,016	758	505	754	749	501	752	503	503

Note 1. Gain – Neutral are the coefficient for those in gain frame when the constant is those respondents in neutral frame. Note 2. These estimations are also controlling for age, gender, self-reported health, marital status, minor children, dependent elderly at home, employment status and education. Note 3. Estimated coefficients above Standard Errors, the latter shown in brackets. Note 4. ***, **, and * mean coefficient is significant at 1%, 5% and 10% of error respectively.

6.3. Effect of reference point of income on MSG responses

In Figures 10 and 11 we see willingness to risk for the prevention different road injuries.[20] In contrast to the findings for CV here we do not find a clear pattern of the effect of the framing. Only for injury X it seems that those with current income equal to normal income are more willing to risk than those in a frame of loss. Also, for S and R there seems to be higher median responses for those respondents in the gain frame than those responses in a neutral situation. The econometric model is shown in Table 13 and is based on the same specification as that of Table 12 for CV although for F, W and X we estimate percentile 70th rather than median responses (see footnote 20). The graphical intuitions are supported by econometric analysis since some framing dummies are significant for X, S and R in the same direction as in the bar graphs. Nonetheless, the most common finding is that framing has no effect on MSG responses since dummies variables are not significant for F, W, V, N and L. With respect to the effect of income (conditional on the framing to be constant) it is generally positive since a positive and significant coefficient is found for those respondents in the interval €2,501-€3,500 (injuries W, V, S, R) and the interval €3,500-€5,000 (injuries X, N, L).

[20] We show median MSG responses except for the three less severe injuries (F, W and X). We show percentile 70th in those cases because median responses are extremely low and do not change among different groups of respondents.

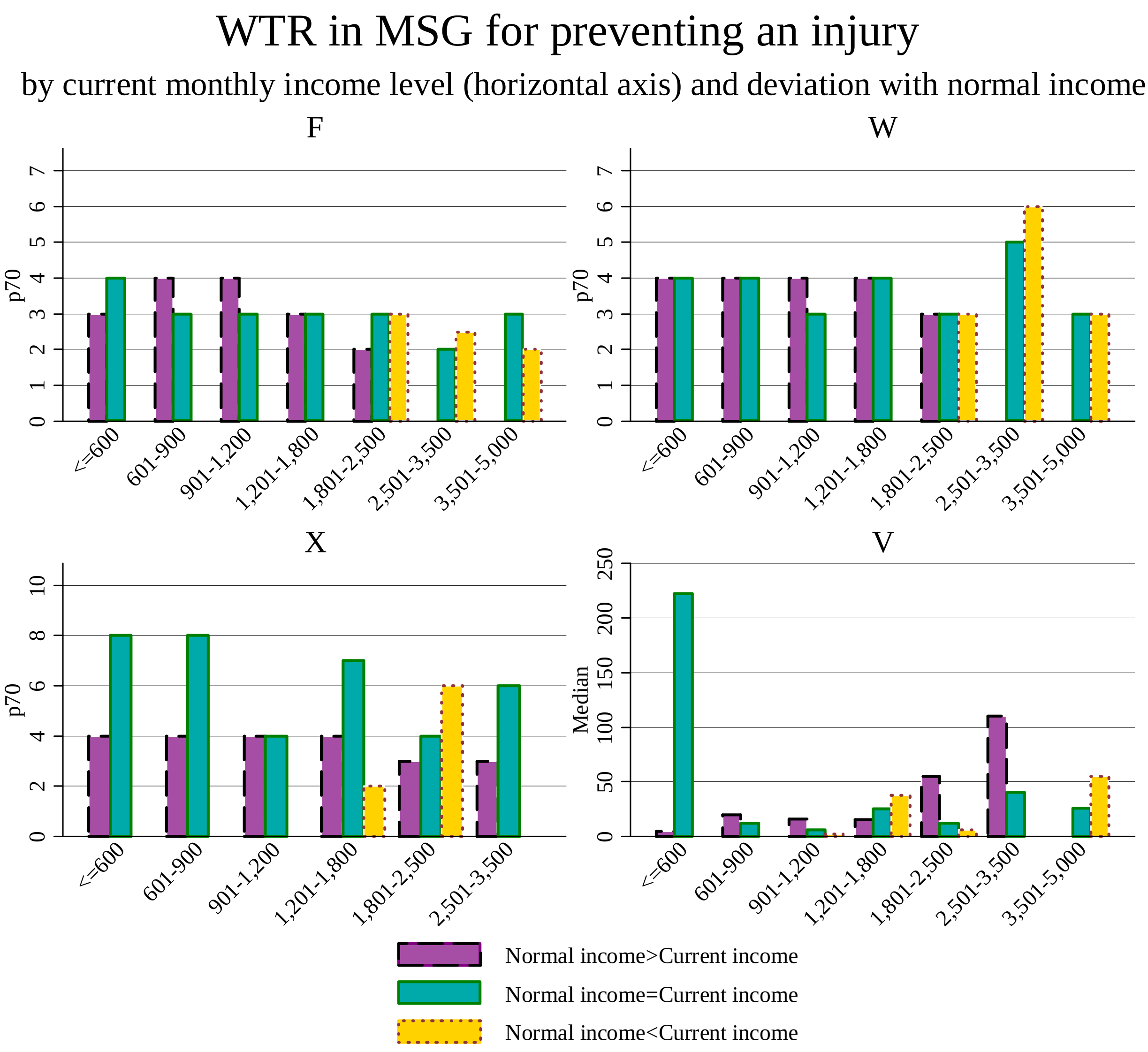


Figure 11. MSG responses (per thousand) by different frames. (Part A)

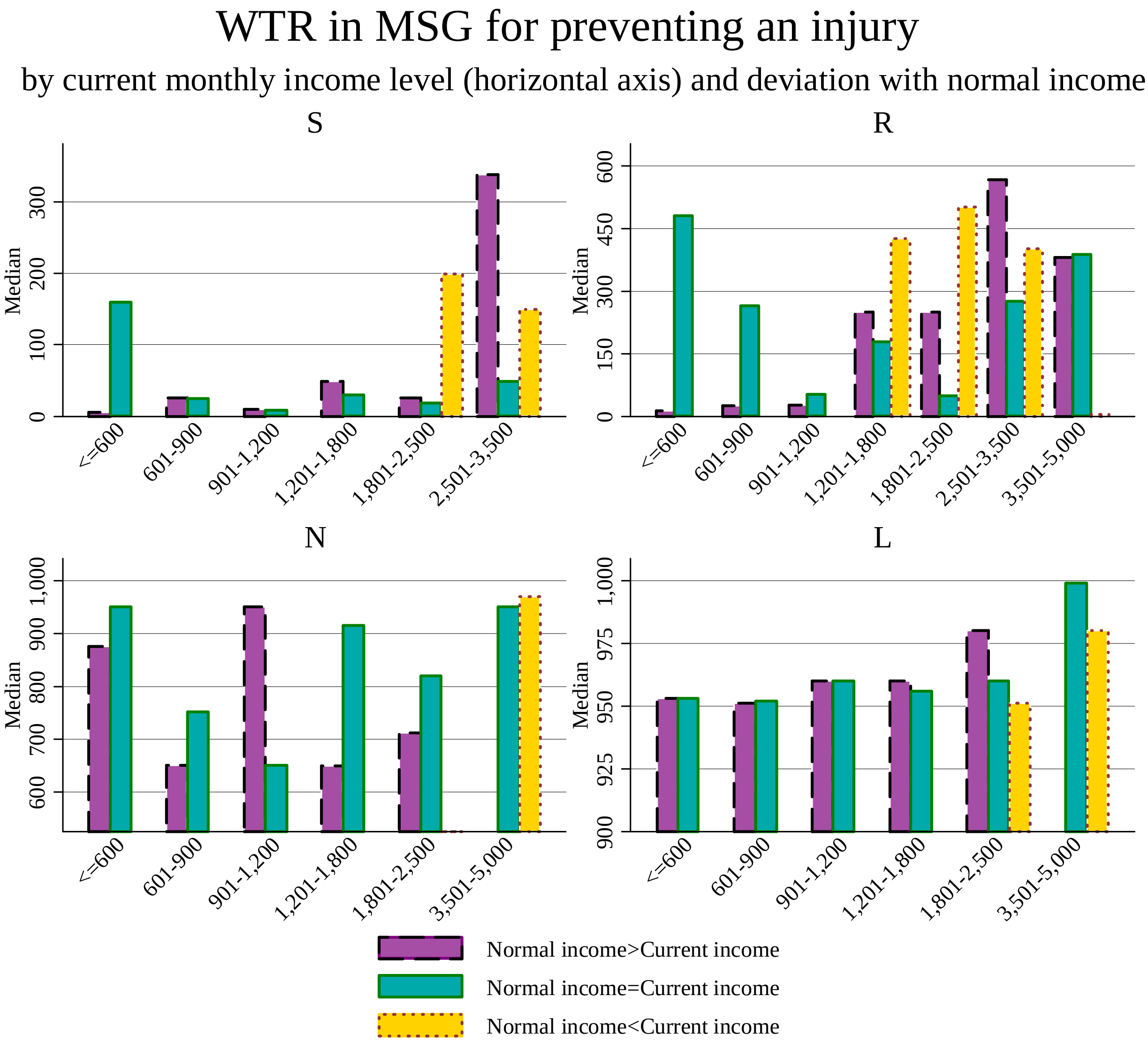


Figure 12. MSG responses (per thousand) by different frames. (Part B)

Table 13. Median and p70 estimation of MSG responses (per thousand)

	MSG response (WTR for preventing:)							
Variables	**F**	**W**	**X**	**V**	**S**	**R**	**N**	**L**
Current Income								
(€, cons: <=900)								
901-1,200	-0.23	-0.02	-0.87	1.07	-6.06	-20.00	31.86	10.24
	(0.31)	(0.33)	(0.57)	(4.80)	(8.55)	(49.56)	(45.16)	(6.55)
1,201-1,800	-0.14	0.26	0.11	8.32	-1.58	52.46	68.38	12.44*
	(0.33)	(0.35)	(0.61)	(5.27)	(9.22)	(53.14)	(49.53)	(7.22)
1,801-2,500	-0.42	0.22	-0.79	9.70	0.13	-6.30	58.70	9.55
	(0.39)	(0.41)	(0.69)	(6.06)	(10.76)	(62.33)	(57.79)	(8.39)
2,501-3,500	-0.31	1.94***	-1.63	32.49***	29.76**	165.89*	97.61	12.44
	(0.53)	(0.56)	(1.00)	(8.54)	(14.82)	(84.86)	(81.09)	(11.88)
>=3,501	0.35	0.09	2.40*	11.52	10.52	62.19	232.01**	29.26*
	(0.74)	(0.76)	(1.32)	(11.36)	(19.31)	(112.56)	(112.03)	(16.35)
Framing								
(Cons: Loss, NI<CI)								
Neutral (NI=CI)	-0.37	-0.14	1.31***	-3.65	3.24	36.64	-32.89	-1.59
	(0.23)	(0.25)	(0.44)	(3.75)	(6.39)	(37.06)	(36.02)	(5.23)
Gain (NI<CI)	-0.75	-0.65	0.55	-8.73	46.31***	206.11**	-132.40	-4.76
	(0.51)	(0.57)	(0.96)	(8.37)	(14.37)	(82.97)	(82.39)	(12.02)
Gain - Neutral	-0.38	-0.51	-0.76	-5.08	43.07***	168.83**	-99.51	-3.17
	(0.48)	(0.54)	(0.90)	(7.90)	(13.61)	(77.51)	(77.64)	(11.35)
Constant	3.70***	3.53**	-0.05	13.21	20.47	167.86	538.35***	957.03***
	(1.31)	(1.39)	(2.43)	(21.18)	(36.91)	(215.25)	(197.42)	(28.55)
Observations	1,011	1,011	1,005	1,005	1,008	1,008	1,008	1,008

Note 1. Gain – Neutral are the coefficient for those in gain frame when the constant is those respondents in neutral frame. Note 2. These estimations are also controlling for age, gender, self-reported health, marital status, minor children, dependent elderly at home, employment status and education. Note 3. Estimated coefficients above Standard Errors, the latter shown in bracket. Note 4. ***, **, and * mean coefficient is significant at 1%, 5% and 10% of error respectively. Note 5. For F, X and W percentile 70th is estimated, the rest of the models are estimating median responses.

6.4. A Reference-Dependent Utility Function to explain framing effects

Surprisingly those with normal income below the current income are willing to pay more for safety improvements than those with normal income above or equal to the current income. We think that this puzzling result can be explain if we consider a reference dependent utility function of income, where the reference point is the permanent income. We can include this feature to the theoretical framework developed in Jones-Lee (1976 and 1989) and Carthy et al. (1999). Under this framework the MRS of wealth for risk of death of an expected utility maximizer individual that faces a risk of death, p, is (see Jones-Lee, 1989; or Chapter 1 of this thesis):

$$m_D = \frac{U(w)-D(w)}{(1-p)U'(w)+pD'(w)}, \tag{5}$$

where $U(w)$ and $D(w)$ are the utility of wealth conditional on normal health and death, respectively, and $U'(w)$ and $D'(w)$ are the corresponding derivatives.

Now we consider that it is the utility function of wealth, w, in normal health, $U(w,r)$, which is reference-dependent, with reference point r. Therefore the theoretical MRS of wealth for risk of death is a modified expression of equation (5):

$$m_D(w,r) = \frac{U(w,r)-D(w)}{(1-p)U_w(w,r)+pD'(w)}. \tag{6}$$

We define here m_D as a function of wealth, w, and the reference point, r. We have a representation of $U(w,r)$ in Figure 13. This function has the following properties:

Property 1. It is an increasing and decreasing function of w and r respectively. So that we have the following:

$$\forall w, U(w,r_1) \lesseqgtr U(w,r_2) \iff r_2 \lesseqgtr r_1$$
$$\forall r, U(w_1,r) \lesseqgtr U(w_2,r) \iff w_1 \lesseqgtr w_2$$

Property 2. Loss aversion. That means that the marginal utility of wealth is higher in the loss frame than in the gain frame. Formally:

$\lambda U_w(w_1,r) = U_w(w_2,r)\ whenever\ w_1 > r > w_2\ and\ w_1 - r = r - w_2$. With $\lambda > 1$.

Property 3. Diminishing sensitivity. The utility function is concave for gains and convex for losses. This is the marginal utility of wealth, $U_w(w,r)$, is decreasing w.r.t. w for gains and increasing for losses. Formally, this property implies that the marginal utility is always higher in the neutral frame (when $w = r$):[21]

$$U_w(w,r) < U_w(w,w)\ whenever\ r \lessgtr w.$$

Property 4. The marginal utility of wealth is a function of the difference between wealth and the reference point. Formally: $U_w(w,r) = f(w-r)$. This property is implicitly satisfied by the *typical* reference-dependent function by assuming that $U(w,r) = g(w-r)$, i.e. the utility of wealth is a function of the difference between wealth and the reference point (for instance Kahneman and Tversky, 1992).[22]

In order to explain the evidence shown in this section we may consider *w* being a representation of current income and *r* being the permanent income. So $U(w,r)$ is a function of current and permanent income, the latter being a reference point. This way we are able to explain income frame effect on $m_D(w,r)$. To do that we will make three further assumptions:

[21] Notice that given the definition of *property 2* (loss aversion) the utility function is not continuously differentiable at the neutral point ($w = r$) so that $U_w(w,w)$ does not exist. However, we are only interested in the left partial derivative of wealth at this point given that we are analysing the change of utility given a reduction in wealth (this is WTP). Therefore it has to be interpreted that $U_w(w,w) = \lim_{w \to r^-} U_w(w,r)$.

[22] Notice that what is assumed by the *typical* reference dependent function is more restricted than *property 4*. For example, in case that wealth and the reference point increase in the same amount (this is relative income remains constant) it would not have an impact on utility. However, *property 4* allows for absolute income to have an impact on utility.

A1. We have r_G, r_N and r_L as the reference points of those in *Gain (G)*, *Neutral (N)* and *Loss (L) scenario* respectively. By definition, permanent income in G (L) is below (above) current income. Also permanent income equals current income in N. Formally, we assume that *given a certain level of w it happens that,* $r_G < r_N = w < r_L$.

A2. *We will also assume* $r_G - w = w - r_L$. This means that the distance between current income and the permanent income is the same for those in the loss and gain frame.[23]

A3. We will assume that the *numerator and denominator in expression (6) are positive* because utility of wealth is higher conditional on normal health than conditional of death, $U(w,r) > D(w)$, and marginal utility is higher than zero, $U_w(w,r) > 0$ and $D'(w) > 0$.

We have the next propositions:

1) *Those respondents in G pay more than those in L, conditional on having the same w.* This is $m_D(w,r_G) > m_D(w,r_L)$. **Proof:** Step 1. Given $A1$ *property 1* implies that $U(w,r_G) > U(w,r_L)$. Step 2. Given $A2$, *property 2* and *property 4* we have $U_w(w,r_G) < U_w(w,r_L)$. To see this imagine w^* such that $w - r_G = r_G - w^*$ then by *property 2* $\lambda U_w(w,r_G) = U_w(w^*,r_G)$. Since we assume $A2$ then we have $r_L - w = r_G - w^*$ and by *property 4* we have $U_w(w,r_L) = U_w(w^*,r_G)$. Therefore $\lambda U_w(w,r_G) = U_w(w,r_L)$. By step 1 and 2 the numerator (denominator) in equation (5) is higher (lower) for those in G. Hence given A3 we have $m_D(w,r_G) > m_D(w,r_L)$■

2) *Those respondents in G pay more than those in N, conditional on having the same w.* This is $m_D(w,r_G) > m_D(w,r_N)$. **Proof:** Step 1. $A1$ and *property 1* imply that $U(w,r_G) > U(w,r_N)$. Step 2. Given $A1$ and *property 3* we know that $U_w(w,r_G) < U_w(w,r_N = w)$. By step 1 and 2 the numerator (denominator) in equation (5) is higher (lower) for those in G. Hence given A3 we have $m_D(w,r_G) > m_D(w,r_N)$■

These two propositions explain the consistent result that those respondents in gain scenario pay more than those in a neutral and loss frame. This is true not only for preventing the risk of a fatality. The same propositions could be proved in the case of WTP for preventing a risk of injury if we consider the theoretical MRS of wealth for risk of injury to be a modified expression of equation (6):

$$m_I(w,r) = \frac{U(w,r) - I(w)}{(1-p)U_w(w,r) + pI'(w)}. \qquad (6)$$

Where the utility of wealth conditional on a road injury, $I(w)$, substitutes the utility conditional on death, $U(w)$.

Notice that we cannot derive a clear result about the difference between those in a neutral frame and those in a loss frame. This is because given *property 1* we have that $U(w,r_N) > U(w,r_L)$ which entails higher willingness to pay for those in neutral frame. However, *property 3* implies $U_w(w,r_L) < U_w(w,r_N)$ which entails the opposite result; higher willingness to pay for those in loss scenario. Therefore, the final results depend on what effect predominates: the utility level effect or the marginal utility effect. This could be the reason why we do not find a clear pattern in the econometric estimations with respect to the neutral and loss frame. For example, we find

[23] We are able to explain the results also if the distance with the reference point is different for *L* and *G*. However for the ease of presentation we see better to make that assumption.

that those in neutral scenario pay less than those in loss scenario for the case of preventing a risk of death and injury S (see Table 12). On the other hand the results are the other way around for the case of injury W. Nonetheless for the majority of the injuries we do not find significant differences between the neutral and loss scenario.

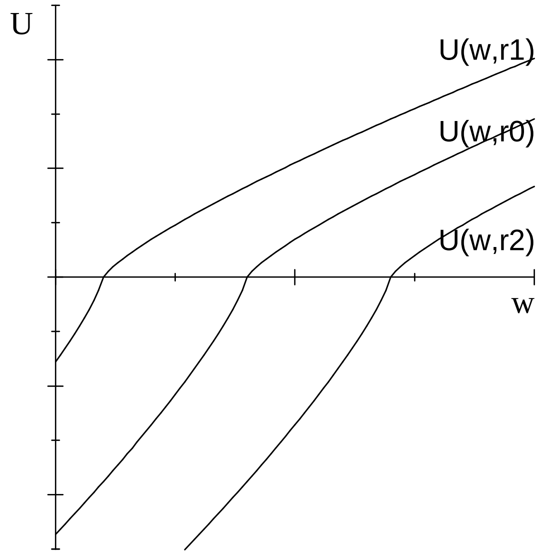


Figure 13. Reference Dependent Utility Functions with r1<r0<r2

Another fact in our analysis is that we find the effect of framing on CV responses to be much more present than on MSG responses. For example, of the 24 framing coefficients in Table 12 ($3\ framing\ effects \times 8\ injuries$) 12 were significant. Only 5 out of 24 were significant in Table 13. To try to understand why this occurs we can analyse the relationship between MSG responses and utility, this is the following (see Carthy et al., 1999; and see Chapter 1 of this thesis):

$$\frac{\pi^{*}-\theta}{1-\theta} = \frac{U(w)-I(w)}{U(w)-D(w)}. \tag{7}$$

However assuming a reference dependent utility function this expression converts into:

$$\frac{\pi^{*}-\theta}{1-\theta} = \frac{U(w,r)-I(w)}{U(w,r)-D(w)}. \tag{8}$$

The theoretical predictions are as follows. Given that for the majority of the injuries it happens that $U(w,w) - I(w) < U(w,w) - D(w)$ then a decrease in r (such as $r < w$) implies (given *property 1*) that $U(w,w) < U(w,r)$ and therefore this gives place to an increment of MSG responses. In other words, those in a frame of gains should be more willing to risk than those in a neutral frame (and also than those in a loss frame). This is what we find in Table 13 for injuries S and R. Also those in a neutral frame should be willing to risk more than those in a loss frame. This is because of *property 1* we have that $U(w,w) > U(w,r)$ when $r > w$ giving place to higher responses in the neutral point. However this theoretical prediction is not by far the evidence encountered because in general MSG responses do not change with the framing.

Given equation (8) we know that theoretically MSG responses only depend on utility but not on the marginal utility of wealth as in the case of CV responses (see equation 6). In other words, MSG responses depend only on *Property 1* but not on the remaining properties as loss aversion and diminishing sensitivity. Therefore one interpretation of the lack of effect of framing on MSG responses is that framing effect is driven mainly by loss aversion and diminishing sensitivity. To consider this as a valid explanation we estimate model of Table 13 but now the dependent variable is the CV relative value of preventing an injury with respect to the value of preventing a fatality rather than raw CV responses. This relative values are computed as the ratio of the MRS of wealth for risk of injury and death:

$$\frac{m_I}{m_D} = \frac{U(w,r)-I(w)}{U(w,r)-D(w)} \cong \frac{wtp_I}{wtp_D} \tag{9}$$

This relative value do not depend on the marginal utility of wealth, actually should be equal to $\frac{\pi^{*}-\theta}{1-\theta}$ (see equation 8).[24] If we find that framing effect on CV relative values are limited our explanation that framing effects found for CV responses in Table 12 are mainly driven by differences in $U_w(w,r)$ between each frame will be consistent. In Table 14 we show coefficients of framing dummies. Only 4 out of 24 framing coefficients are statistically significant therefore our interpretation makes sense.

[24] This theoretical relation between CV relative values and MSG responses implies that there are two ways of eliciting this relative value, with CV and MSG. However evidence shows that CV relative values differ from MSG ones (see Jones-Lee et al., 1995; and Chapter 1 of this thesis). Nevertheless the interesting fact here is that both relative values do not depend on $U_w(w,r)$.

Table 14. Median and p80 estimation of CV relative values

Variables	Relative value of preventing: F	W	X	V	S	R	N	L
Framing								
(Cons: Loss, NI<CI)								
Neutral (NI=CI)	0.08**	0.12**	-0.002	-0.02	0.13***	0.04	-0.07	-0.20
	(0.04)	(0.06)	(0.03)	(0.05)	(0.04)	(0.04)	(0.27)	(0.17)
Gain (NI<CI)	0.06	0.08	0.004	-0.04	-0.04	0.08	-0.02	-0.30
	(0.08)	(0.13)	(0.08)	(0.10)	(0.08)	(0.08)	(0.70)	(0.48)
Gain – Neutral	-0.03	-0.04	0.00	-0.02	-0.16**	0.03	0.04	-0.10
	(0.08)	(0.12)	(0.07)	(0.13)	(0.08)	(0.07)	(0.66)	(0.45)
Constant	0.39*	0.36	0.99***	0.91**	1.77***	1.12***	0.84	1.88**
	(0.20)	(0.34)	(0.19)	(0.39)	(0.24)	(0.22)	(1.51)	(0.86)
Observations	665	435	683	685	454	661	431	431

Note 1. Gain – Neutral are the coefficient for those in gain frame when the constant is those respondents in neutral frame. Note 2. These estimations are also controlling for current income, age, gender, self-reported health, marital status, minor children, dependent elderly at home, employment status and education. Note 3. Estimated coefficients above Standard Errors, the latter shown in brackets. Note 4. ***, **, and * mean coefficient is significant at 1%, 5% and 10% of error respectively. Note 5. For N and L percentile 80th is estimated, the rest of the models are estimating median responses.

6.5. Discussion

The analysis suggests that there is a relative income component affecting the life satisfaction and monetary valuation of road safety. This finding is interesting because it allows more precise predictions about WTP. In addition to the current income, the permanent income plays a role on predicting WTP for safety improvements. Even more, its effect is not a positive effect on WTP as the permanent income hypothesis and theory would suggest. Its role is more of a reference point or income aspiration. Specifically, given a specific current income those with a lower permanent income (those in gain scenario) are willing to pay more than those with lower or equal permanent income (those in loss and neutral scenario). This phenomenon is coherent with a reference-dependent representation of utility as proposed in previous section. This utility function is in line with the idea of Clark et al. 2008 (p. 128): "*…at a given level of own current income yt, an individual whose income has just increased* (gain frame) *has higher utility than someone whose income has just decreased* (loss frame)*…* .[25]

From a normative perspective there are implications. The relative evaluation of income when deciding the consumption of safety seems to be a psychological bias. There are no obvious normative reasons for supporting the fact that a person should buy a safety device just because she/he is in a stage of high earnings in her/his life cycle, even more if those with the same current income in a low earnings stage would not buy it. In other words, it seems a mistake that those who pay more (those in gain frame) have a lower permanent income than those who pay less (those in neutral and loss scenario).

Since VSL and VPI are based on individual valuation of safety there could be a bias in aggregate values of preventing a fatality or a road injury. For example, in those countries in economic recession VSL will be lower than in those countries in the expansionary phase, even though current National Income is the same. The immediate implication is that VPL and VPI

[25] We add text in parentheses for explanation purposes.

should be downward corrected for those countries in economic expansion with respect to those in economic recession. There are also implications for markets in which people exchange money for safety. Financial advisors have grounds to check whether their clients are affected by their income reference point and help them make better decisions when buying a safety device for a car or a health insurance for instance. Future studies could analyse this issue. Eventually people may not want to change their decisions even when they are aware about their biases. If this is not the case the justification for bias correction is stronger.

On the other hand, there are reasons to be cautious in this matter as well. For example, misprediction of future income could have a role here. Imagine that those who have just experienced an increment in their current income, those in gain scenario, do not realize that part of that increment is permanent. Although their aspiration/reference point would not change sharply, so they are in a gain scenario, they would be underestimating their permanent income. Eventually this would be a normative justification for them to expend more on safety. The study of different biases that could play a role may shed light about bias correction adequacy.

One further comment can be said about future research on relative income and safety valuation. First, the permanent income that we consider as a reference point is directly reported by the respondents. They are supposed to report an average "normal income" throughout their entire lives considering past and (expected) future income. This entails a twofold implication: on the one hand no assumptions have to be made about the reference point; on the other hand it is a subjective measure. The former implication should mean that we capture the individual reference/aspiration better than when it is approximated with past income as in previous works (Clark, 1999; Weinzierl, 2005; Grund and Sliwka, 2007). The latter implies that reported permanent income may be correlated with unobservable personality traits and estimations may be biased.

Eventually we arise again to the same conclusion that MSG is more invariant to different characteristic of the valuation. In this case we find the effect of the income reference to be present in CV responses but not in MSG. Therefore MSG is more in line with normative theory that predict that valuation of injuries should not depend on reference points. In this sense MSG seems to be a more appropriate method for eliciting preferences over safety. However, MSG can only be used to elicit the relative value and not monetary values of preventing an injury. At the end of the day if we want to perform cost-benefit analysis some economic limit or boundary has to be put for public investment on road safety. In this sense the results found in this study shed light about what type of biases can be found for this method and correct responses consequently.

7. Conclusion

Valuing the prevention of fatal and Non-Fatal Injury is crucial for implementing an assessment of traffic safety programs. State preference methods are considered as the suitable procedure for this task. Given the differences encountered in previous studies between the two major procedures used so far in the literature, CV and SG, it is important to understand why this is so and what other features could be taken into account to test the relative performance of them. This chapter provides an analysis of three "framing" effects that can be affecting the performance of these methods. First we analyse the differences between Separate and Joint

Evaluation of road safety. Then we test differences between JE contexts. Finally, we analyse the willingness to pay and willingness to risk in different frames depending on the presumed frame that respondent are located in according to their income reference point: loss, neutral or gain. As a general conclusion we find that all these context effects seems to be presented in the case of CV, however the method of MSG is less influenced by any of them.

With respect to the first analysis we conclude that the evaluation of NFRIs is more value sensitive in JE mode for both methods MSG and CV. However, value sensitivity is largely decreased in SE for CV but only slightly decreased for MSG. We find a systematic JE-SE bias: mildest HSs have a higher value in SE and most severe HSs have a higher value in JE. This results leads to at least two implications. On the one hand, not only CV has been demonstrated to be a method that gives place to insensitive responses to the severity of injuries in previous studies but that its insensitivity is even higher if the evaluation is provided in SE. And on the other hand, if evaluation of safety in the Joint context is considered as a more proper Evaluation Mode then there is a way to correct SE valuations of NFRIs, this is by incrementing (decreasing) the value of mild (serious) injuries.

The second analysis does not find clear context effects in JE. We do not find that the responses differences between contexts are consistent with the range principle effect of Range-Frequency Theory. However we do find evidence of frequency effects for CV. Eventually, the predicted anchor effects were not found. The main conclusion of this analysis is that MSG is proved to be a more consistent method since valuation does not changes among different contexts as much as valuation by CV does.

For the eventual analysis shows that people evaluate their life satisfaction considering reference points. If they are above that reference they considered themselves happier than even other individuals that with the same income are below their reference point. Moreover, this reference dependent evaluation of income has an actual impact on the evaluation of safety by CV. The effect on CV responses of being in different frames (loss, neutral or gain) seems to be consistent with the properties of a reference-dependent utility function of income. What is interesting is that the framing effect seems to be far less important for the MSG method which can be explained by the fact that these framing effects are mainly driven by changes in the marginal utility of wealth in the different frames. Neither the CV relative values of preventing an injury with respect to the value of avoiding a fatality seem to be affected by framing effect which turns to be another good reason for using this ratio for policy purposes rather than absolute value of preventing an injury.

References

1. Abellán, J.M., Martínez, J. E., Méndez, I., Pinto, J.L., Sánchez, I. (2011). *El valor monetario de una vida estadística en España. Estimación en el contexto de los accidentes de tráfico. Dirección General de Tráfico.*
2. Ariely, D., Loewenstein, G., and Prelec, D., (2003). “Coherent Arbitrariness”: stable demand curves without stable preferences. *The Quarterly Journal of Economics* 118, 73–105.
3. Atkinson, S.E., and Halvorson, R. (1990). The valuation of risks to life: evidence from the market for automobiles. *Review of Economics and Statistics,* 73, 133–136.
4. Andersson, H. (2005). The value of safety as revealed in the Swedish car market: An application of the hedonic pricing approach. Journal of Risk and Uncertainty, 30(3), 211-239.
5. Bartels, D. M. (2006). Proportion dominance: The generality and variability of favoring relative savings over absolute savings. *Organizational Behavior and Human Decision Processes, 100*(1), 76-95.
6. Bateman, I. J., Cameron, M. P., and Tsoumas, A. (2007). Investigating the characteristics of stated preferences for reducing the impacts of air pollution. *Environmental Economics, Experimental Methods, 10*, 424.
7. Blomquist, G.C. (1979). Value of life saving: implications of consumption activity. *Journal of Political Economy,* 78, 540–558.
8. Blomquist, G. C., Miller, T. R., and Levy, D. T. (1996). Values of risk reduction implied by motorist use of protection equipment: New evidence from different populations. *Journal of Transport Economics and Policy*, 30(1), 55-66.
9. Carthy, T., Chilton, S., Covey, D., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N., and Spencer, A. (1999). On the contingent valuation of safety and the safety of contingent valuation: Part 2 - the CV/SG "chained" approach. *Journal of Risk and Uncertainty, 17*(3), 187-213.
10. Cervone, D., and Peake, P. K., (1986). Anchoring, efficacy, and action: the influence of judgmental heuristics on self-efficacy judgments and behavior. *Journal of Personality and Social Psychology*, 50, 492–501.
11. Chapman, G. B., and Johnson, E. J., (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes*, 79, 1–39.
12. Chu, P. Y., Chang, C. C., Chen, C. Y., and Wang, T. Y. (2010). Countering negative country-of-origin effects: The role of evaluation mode. *European Journal of Marketing*, *44*(7/8), 1055-1076.
13. Clark, A. E. (1999). Are Wages Habit-Forming? Evidence from Micro Data. *Journal of Economic Behavior and Organization*, 39(2), 179–200.
14. Clark, A. E., Frijters, P., and Shields, M. (2008). Relative income, happiness and utility: An explanation for the Easterlin paradox and other puzzles. *Journal of Economic Literature*, 46, 95–144.
15. Covey, J., Robinson, A., Jones-Lee, M., and Loomes, G. (2010). Responsibility, scale and the valuation of rail safety. *Journal of Risk and Uncertainty, 40*(1), 85-108.

16. Critcher, C. R., and Gilovich, T., (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making* 21, 241–251.

17. De Blaeij, A. T., Rietveld, P., Verhoef, E. T., and Nijkamp, P., (2002). The valuation of a statistical life in road safety: a stated preference approach, *30th European Transport Forum*. London, UK.

18. Englich, B., Soder, K., (2009). Moody experts – how mood and expertise influence judgmental anchoring. *Judgmental and Decision Making* 4, 41–50.

19. Epley, N., and Gilovich, T. (2001). Putting adjustment back into the anchoring and adjustment heuristic: differential processing of self-generated and experimenter-provided anchors. *Psychological Science* 12, 391–396.

20. Easterlin, R. A. (2001). Income and happiness: Towards a unified theory. *The Economic Journal*, *111*(473): 465-484.

21. Easterlin, R.A. (1995). Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior and Organization*, 27, 35–47.

22. Easterlin, R.A. (1974). Does economic growth improve the human lot? In P.A. David and M.W. Reder (Eds.), Nations and households in economic growth: Essays in honor of Moses Abramovitz (pp. 89–125). New York: Academic Press.

23. Friedman, M. (1963). Windfalls, the Horizon, and Related Concepts in the Permanent Income Hypothesis. *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, edited by Carl Christ et al. Standford, Calif.: Standford Univ. Press.

24. Friedman, M. (1957). A Theory of the Consumption Function. Princeton, N.J.: Princeton Univ. Press.

25. Furnham, A., and Boo, H. C. (2011). A literature review of the anchoring effect. *Journal of Socio-Economics*, *40*(1), 35-42.

26. Galinsky, A. D., and Mussweiler, T., (2001). First offers as anchors: the role of perspective taking and negotiator focus. *Journal of Personality and Social Psychology*, 81, 657–669.

27. Grund, C., and Sliwka, D. (2007). Reference-Dependent Preferences and the Impact of Wage Increases on Job Satisfaction: Theory and Evidence. *Journal of Institutional and Theoretical Economics*,163(2): 313–35.

28. Hammitt, J. K., and Haninger, K. (2010). Valuing fatal risks to children and adults: Effects of disease, latency, and risk aversion. *Journal of Risk and Uncertainty*, *40*(1), 57-83.

29. Hsee, C. K (1996). The evaluability hypothesis: An explanation of preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *46*, 247-257.

30. Hsee, C. K., Blount, S., Loewenstein, G. F., and Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576-590.

31. Hsee, C. K., Yang, Y., Li, N., and Shen, L. (2009). Wealth, warmth, and well-being: Whether happiness is relative or absolute depends on whether it is about money, acquisition, or consumption. *Journal of Marketing Research*, 46, 396–409.

32. Hsee, C. K., and Zhang, J. (2010). General Evaluability Theory. *Perspectives on Psychological Science*, *5*(4), 343-355.

33. Jones-Lee, M., Loomes, G., and Philips, P. (1995). Valuing the prevention of nonfatal road injuries - contingent valuation vs standard gambles. *Oxford Economic Papers-New Series, 47*(4), 676.

34. Jones-Lee M. (1989). The economics of safety and physical risk. *Oxford: Basil Blackwell*.

35. Jones-Lee, M. (1985). The Value of Safety: Results of a National Sample Survey. *The Economic Journal*, 95, 49-72.

36. Jones-Lee, M. (1976). The Value of a Statistical Life: An Economic Analysis. *University of Chicago Press*, Chicago.

37. Kahneman, D. (2008) The sad tale of the aspiration treadmill, http://www.edge.org/q2008/q08_17.html

38. Kahneman, D., and Tversky, A. (1979). Prospect theory - analysis of decision under risk. *Econometrica, 47*(2), 263-291.

39. Kniesner, T. J., Viscusi, W. K., Woock, C., and Ziliak, J. P. (2012). The value of a statistical life: Evidence from panel data. *Review of Economics and Statistics, 94*(1), 74-87.

40. Kniesner, T. J., and Viscusi, W. K. (2005). Value of a statistical life: Relative position vs. relative age. *American Economic Review, 95*(2), 142-146.

41. Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* 46, 33–50.

42. Lichtenstein, S., and Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89*(1), 46-55.

43. Melinek, S. J. (1974). A method of evaluating human life for economic purposes. *Accident Analysis and Prevention,* 6, 103–114.

44. Parducci, A. (1965). Category judgement: a Range-Frequency model. *Psychological Review,* 72(6), 407-418.

45. Persson, U., Trawén, A., Norinder, A., Hjalte, K., and Andersson, H. (2001a). Relative Risk Values of Non-Fatal Traffic Injuries - A Comparison Between Contingent Valuation, Risk-Risk Trade Off and Standard Gamble Methods. *Swiss Journal of Economics And Statistics*, 137 (1), 117-128.

46. Persson, U., Norinder, A., Hjalte, K., Gralén, K. (2001b). The Value of a Statistical Life in Transport: Findings from a New Contingent Valuation Study in Sweden. *Journal of Risk and Uncertainty,* 23(2), 121-134.

47. Persson, U., and Cedervall, M. (1991). The value of risk reduction: results of a Swedish sample survey. *The Swedish Institute for Health Economics, Lund, IHE Working Paper* 1991:6, Lund, Sweden.

48. Persson, U., Norinder, A., and Svensen, M. (1995). Valuing the benefits of reducing the risk of non-fatal road injuries: the Swedish experiment. In: Schwab Christe, N.G., Soguel, N.C. (Eds.), *Contingent Valuation, Transport Safety and the Value of Lif*e. Kluwer Academic Publishers, Boston, 63–84.

49. Robinson, A., Loomes, G., Jones-Lee, M. (2001). Visual analog scales, standard gambles, and relative risk aversion. *Medical Decision Making,* 21(1), 17–27.

50. Spackman, M., Evans, A., Jones-Lee, M., Loomes, G., Holder, S., Webb, H., and Sugden, R. (2011). Updating the VPF and VPIs: Phase 1: Final Report Department for Transport. London: NERA Economic Consulting.

51. Stalmeier, P. F. M. (2002). Discrepancies between chained and classic utilities induced by anchoring with occasional adjustments. *Medical Decision Making, 22*(1), 53-64.

52. Stevenson, B., and Wolfers, J. (2008). Economic growth and subjective well-being: Reassessing the Easterlin paradox. Bookings Papers on Economic Activity, 1–87.

53. Tversky, A., and Kahneman, D. (1992). Advances in prospect-theory - cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297-323.

54. Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Biases in judgments reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124-1131.

55. Viscusi, W. K., and Huber, J. (2012). Reference-dependent valuations of risk: Why willingness-to-accept exceeds willingness-to-pay. *Journal of Risk and Uncertainty, 44*(1), 19-44.

56. Viscusi, W. K. (2010). The heterogeneity of the value of statistical life: Introduction and overview. *Journal of Risk and Uncertainty*, 40(1), 1-13.

57. Viscusi, W. K., and Aldy, J. E. (2003). The value of a statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty*, 27(1), 5-76.

58. Wedell, D. H., Parducci, A. (1988). The category effect in social judgment: experimental ratings of happiness. *Journal of Personality and Social Psychology*, 55(3); 341-356.

59. Weinzierl, M. (2005). Estimating a Relative Utility Function. *Harvard University*.

60. Winston, C., and Mannering, F. (1984). Consumer demand for automobile safety: new evidence on the demand for safety and the behavioral response to safety regulation. American Economic Review 72, 316–31.

Chapter 3. A range-frequency theory application to context effects in the health domain

1. Background

A relevant issue in economic analysis is whether context, defined by the set of alternatives available to a decision maker, affects decisions that are supposed to be founded in utilities attached to each alternative. Marketing research shows that choices made by consumers are affected by the context in which the decision is made. For example, Simonson and Tversky (1992) and Tversky and Simonson (1993) found that preferences for an option *x* over *y* can reverse when a third option *z* is added to the choice set or when the background context (alternatives seen before the decision is made) is changed. Specifically in the present study we are concerned with context effects in the health domain. In this sense there are some important questions: ¿Are health states utilities context-dependent?; ¿What theoretical model is a good account of this phenomenon?; and if utility elicitation does vary with context ¿What utility estimation should be used for health programs evaluation? ¿Is it possible to correct for context effects in order to elicit the so called "context-free" health utilities?

In order to answer these questions it is convenient to consider research done in the psychophysical empirical literature. There exists a regular empirical finding in experimental settings that the context in which an object is evaluated affect to the eventual judgment attached to it. For example, a square looks big (small) when presented jointly with small (big) squares. This is what is known as a contrast effect. A good psychological account of this phenomenon is Parducci's (1965) Range Frequency Theory (RFT), in which an object is assigned a value (rate or category) depending on two principles, namely the *range* and *frequency*. For example, in size judgments this model predicts that an object like a square will be judge as "larger" in a context where the range and/or the frequency value for that object is higher. The range value refers to the proportion of the contextual psychological range that is below the subjective value of the object. More specifically, given the subjective value of the target object, S_i, and of the two extreme objects in the context (i.e. the largest and smallest squares), S_{min} and S_{max}, the range value of i in a context *k* is computed as $R_{i,k} = \frac{S_i - S_{min,k}}{S_{max,k} - S_{min,k}}$. On the other hand, the frequency value represents the proportion of number of stimuli with subjective value below the target stimulus in the context, computed as $F_{i,k} = \frac{Rank_{i,k} - 1}{N_k - 1}$, where $Rank_{i,k}$ is the rank of the target stimulus in the context and N_k is the number of stimuli in the context. The contextual judgment is a compromise of the two principles and is computed as $J_{i,k} = \omega R_{i,k} + (1 - \omega) F_{i,k}$, where $\omega \in [0,1]$ is a weighting parameter measuring the relative importance of the two psychological principles involved in contextual valuations (see Wedell and Parducci, 1988, for the algebraic representation of RFT).

Notice that the two principles of RFT are consistent with the logic of contrast effects. For example, in the case of squares a high $R_{i,k}$ means that square *i* is closer to the largest square included in context *k* and far from the smallest one, this square would be judged as a large square because it looks large in comparison to the psychological range included in context *k*. Experimental manipulation of $S_{min,k}$ and $S_{max,k}$ would derive in contrast effects of this type. Also a high $F_{i,k}$ implies that in context *k* there is a high proportion of squares that are smaller than square *i* implying a high judgment of size in comparison to the rest of the squares included in the context. Manipulation of the number of squares included in the context, N_k, and/or the rank of the target square would lead to contrast effects.

Another model of contrast is given by Adaptation Level Theory (AL by Helson, 1964) that entails that contextual valuation of a stimuli is given by $J_{i,k} = a + b(S_i - AL_k)$ where a and b are scaling constants, S_i has the same interpretation as in RFT and AL_k is the adaptation level that can be interpreted as a comprehensive reference point in the evaluation context. For example, AL predicts that evaluation would be higher in a context where the average subjective value of contextual stimuli is lower.

Nonetheless, it has been proved that RFT is a better account than AL of context effects in psychophysical domains like the judgment of square size (Parducci, 1965; Birnbaum, 1974 and 1999; Parducci and Wedell, 1986; Wedell, 2008), body size (Wedell et al., 2005), tempo of music (Rashotte and Wedell, 2012). However the model explain context effects found in social domains like rating of happiness (Wedell and Parducci, 1988), assignment of grades to students (Wedell et al., 1989), employees wage satisfaction rating (Brown et al., 2008); or physiological domains like evaluation of pain (Watkinson et al., 2013).

In the health domain previous studies have found context effects related to manipulation of ranks (Bleichrodt and Johannesson, 1997a; Robinson et al., 2001). For example, Bleichrodt and Johannesson (1997a) found that the rate assigned to a health state on a Visual Analogue Scale (VAS) depended on the number of preferred alternatives in the context of evaluation. Specifically, four health states (A, B, C and D) were common to two experimental groups. Ranks were manipulated for one of them (B) being rated higher in the group where it was the third best health state rather than in the group where it was the fourth best. However, no context effects were found for the remaining common health states which had the same number of preferred health states in both contexts. This result is consistent with range-frequency theory and turn to be the main reason for supporting the idea of using that model to correct for context effect bias in order to estimate what has been called "context-free" valuation of heath states. Schwartz (1998) proposed to estimate true underlying preferences making used of the context information in which health states were valued once the process leading to context effects is known (e.g. we know that range frequency is an account of these context effects). For this purpose there must exist an invariant context-free value attached to each health state that correspond with S_i (i.e the subjective value in RFT). It is assume that this underlying value combined with the context information lead to the contextual judgment. Separating the context information from the true underlying value could be possible when the same health state is valued in two different contexts. Robinson et al. (2001) applied the same reasoning to estimate underlying context free health VAS values. In their analysis respondents gave a higher rating to the same health state when this had a higher rank, however when applying range frequency theory to elicit true underlying values differences between contexts disappeared. This is evidence for both the necessity and suitability of RFT to control for VAS context effects. On the other hand, they did not found significant context effects for utilities elicited by Standard Gamble (SG).

In general there is a lack of evidence making clear if utility elicitation techniques used in Health Economics, like SG or Time Trade Off (TTO), are subjected to RFT effects. Neither there is evidence investigating whether it is possible to compute context-corrected utilities that should be used for policy evaluation purposes. Nonetheless, recently Pinto (2013) presented evidence that suggests that TTO are affected by ordering effects consistent with RFT. Specifically he found that the valuation of a health state is negatively affected by the severity of the preceding health situation evaluated in the context. Here the analysis of an experiment is presented in which context effects for TTO are studied. This analysis is relevant given that TTO is one major procedure for utility elicitation (Torrance, 2005). In this framework health profiles are defined by two attributes, namely quality of life (Q) and life expectancy (T). Individuals have to decide how much T is willing to give up in order to increase Q, specifically respondents trade life years for a change in quality of life from bad to Full Health. In this sense valuation of a health profile is inversely related to the sacrificed life years. The frequency value of target health profiles is manipulated in two separate ways: 1) Five different groups are constructed in such a way that quality of life associated to the contextual health profiles and the number of health profiles included in the context varies among them; and 2) additional two groups differ in life expectancy for the health profiles included in the context. The experimental design allows us to test for: a) basic contrast effects predicted by RFT and other contrast effect theories like Adaptation Level Theory; b) More elaborated RFT effects not predicted by AL; c) Context effects not predicted by RFT.

Results suggest that RFT predicts properly the context effects, due to manipulations of both Q and T, while AL fails to predict some specific effects. With few exceptions between-groups differences follow differences in frequency value, i.e. the higher is the rank of a health profile the less life expectancy an individual is willing to give up. However fitting the RFT model to data does not derive in invariant “context-free” utilities. Specifically, these context-free utilities are the same for groups with the same number of health states in the context but they differ between groups with different number of contextual stimuli.

In the next we show the theoretical framework for RFT and computation of context free utilities of quality of life. Section 3 describes the details of the survey and experimental design. In section 4 results are presented and section 5 contains a discussion. Finally, section 6 is the conclusion.

2. Theory

A health profile (Q_i, T_i) is defined by two attributes, quality of life or health state (Q_i) and life years (T_i). In the QALY (Quality Adjusted Life Year) model its value is computed as:

$$QALY(Q_i, T_i) = Q_i \times D(T_i). \tag{1}$$

Where $D(T_i)$ are the discounted life years. A Time Trade-off (TTO) exercise consists on finding a number of years in Full Health (T_{FH}) that are indifferent to (Q_i, T_i):

$$Q_i \times D(T_i) = Q_{FH} \times D(T_{FH}). \tag{2}$$

Where Q_{FH} is the quality of life in Full Health that we can normalize to unity $(Q_{FH} = 1)$. T_{FH} may be considered as a measure of the valuation of (Q_i, T_i). Now we can compute the “context-free” quality of life or context free utility of health state *i*:

$$Q_i = \frac{D(T_{FH})}{D(T_i)}. \tag{3}$$

However in the present study we assume that the context affects the valuation of (Q_i, T_i). This is the number of years in Full Health indifferent to (Q_i, T_i) changes with the context. Therefore the quality of life or utility of *i* in context *k* is computed as:

$$Q_{i,k} = \frac{D(T_{FH,k})}{D(T_i)}. \tag{4}$$

We will call $Q_{i,k}$ the context dependent utility which is directly given by individual valuation in context k (i.e. by $T_{FH,k}$). However Q_i is unobservable and have to be inferred by applying RFT. Now we can apply two different approaches to RFT. The first one considers that it is the quality of life that is affected by its range value and frequency value in context *k*.

$$Q_{i,k} = \omega \times \frac{Q_i - Q_{min,k}}{Q_{max,k} - Q_{min,k}} + (1-\omega) \times \frac{Rank_{i,k} - 1}{N_k - 1}. \tag{5}$$

Where ω is the weighting parameter for the range/frequency principle. $Q_{min,k}$ and $Q_{max,k}$ are the lowest and highest, respectively, quality of life in context k and affect to the range value $R_{i,k} = \frac{Q_i - Q_{min,k}}{Q_{max,k} - Q_{min,k}}$. In our study we keep the range value constant across contexts with $Q_{min,k} = 0$ as "quality of life" in Death and $Q_{max,k} = 1$ as quality of life in Full Health. $Rank_{i,k}$ refers to the rank of Q_i and N_k to the number of qualities of life included in *k*. Both variables form the frequency value, $F_{i,k} = \frac{Rank_{i,k} - 1}{N_k - 1}$. Model in expression (5) is called Rank Dependent Valuation Model (RDV) and as Wedell (1995) says: "*The rank-dependent valuation model is an instantiation of range-frequency theory in which rank-order effects operate at the valuation stage rather than at a response selection stage in the judgment process.* We will use this model to construct context free utilities when experimental manipulation of contexts is made at the quality dimension.

On the other hand we can apply a more classical or original RFT model that operates at the response selection stage. In this case the valuation of (Q_i, T_i) expressed by the response $T_{FH,k}$ depends on its range value and frequency value rescaled to the response variable scale:

$$T_{FH,k} = \left[\omega \times \frac{T_{FH} - T_{min,k}}{T_{max,k} - T_{min,k}} + (1-\omega) \times \frac{Rank_{i,k} - 1}{N_k - 1}\right]\left[T_{max,k} - T_{min,k}\right] + T_{min,k}. \tag{6}$$

Where $T_{min,k}$ and $T_{max,k}$ are responses associated to the worst and best Health Profiles included in context *k*, respectively. Since in our experimental design we include Death and $(Q_{FH}, 20)$ as the worst and best health profiles in every context we keep them constant with $T_{min,k} = 0$ and $T_{max,k} = 20$. Now $Rank_{i,k}$ and N_k refer to the rank of (Q_i, T_i) and number of health profiles included in *k*, respectively. We will refer to model in expression (6) as R-RFT and we will use it to construct context free utilities when we experimentally manipulate the *life years* dimension and maintain quality of life constant across contexts.

We are really interested in the true underlying utility Q_i. If we combine (4) and (5) this context free quality of life is computed under RDV as (see appendix for algebraic operations):

$$Q_i = \frac{D(T_{FH,k})}{D(T_i)\omega} - \frac{(1-\omega)}{\omega} \times F_{i,k}. \tag{7}$$

When we apply R-RFT at the response stage the context free utility is computed as (see appendix):

$$Q_i = \frac{D\left(\left[\frac{T_{FH,k}}{20\omega} - \frac{(1-\omega)}{\omega} \times F_{i,k}\right] \times 20\right)}{D(T_i)}. \tag{8}$$

Notice that predictions of RDV and R-RFT are similar in the sense that valuation of a health profile (Q_i, T_i) is higher in those contexts where its frequency value is higher. The frequency value of a health profile can be manipulated by changing the quality of life Q_i associated to the contextual health profiles. For example, (Q_i, T_i) is seen as a better health profile when it is evaluated in a context with other health profiles (Q_j, T_j) $j = 1,..,J$ such that $Q_j < Q_i$ and $T_j = T_i$ $\forall j = 1,..,J$. The frequency value can also be manipulated by changing the life years of the contextual health profiles. For example, (Q_i, T_i) is considered as a better health profile in a context where other health profiles are (Q_j, T_j) $j = 1,..,J$ such that $Q_j = Q_i$ and $T_j < T_i$ $\forall j = 1,..,J$. However, we see more appropriate to use RDV when manipulation occurs solely at the quality dimension and R-RFT in other cases, like when we manipulate life years of the contextual health profiles.

3. Survey

3.1. Respondents

A total of 587 individuals participated in an online survey. They were invited to participate in a survey aimed at studying decisions and valuations in the health domain by university researchers. A consulting firm was in charge for the collection of data and recruitment of participants from a set of Spanish population familiar with previous online surveys. Respondents were paid a fixed amount of bonuses for their participation that they could exchange for other items. Characteristics of subjects can be found in Table 1. As we see half of respondents were female with age ranging from 18 to 69. Most of them had secondary or tertiary education. With respect to employment status the majority of them were workers or inactive people.

Table 1. Distribution of respondents by characteristics.

Variables	%	Variables	%
Gender:		**Education:**	
Male	49.6	Primary	7.5
Female	50.4	Secondary	48.9
		Tertiary	43.6
Age:			
18-25	10.1	**Employment Status:**	
26-35	22.2	Worker	43.4
36-45	25.2	Self Employed	14.9
46-55	22.2	Unemployed	18.1
56-69	20.4	Inactive	23.5

3.2. Manipulation of contexts

Quality of life associated to each health prolife was determined by three health related attributes included in the Health Utilities Index Mark 3 (HUI3; Feeny et al., 1995): ambulation; pain and emotion. The descriptions of levels used in the HUI3 for the attributes are described in Table 2; the higher is the level the lower the quality of life. In the survey we translate descriptions to Spanish following Ruiz et al. (2013). We describe each quality of life by the levels of each attribute. For example Q332 represents ambulation and pain with level 3 and emotion with level 2. Finally a health profile is described as the number of years associated to a quality of life. For example, we use (Q332, 20) for 20 years in Q332.

An evaluation context is characterised by the health profiles that are included to be valued. Health profiles, their frequency values and number of subjects for each context are shown in Table 3. In five contexts (A, B, C, D, E) all health profiles had the same duration (20 years) and quality of life was manipulated. Also, number of health profiles was changed in order to see whether manipulation of frequency values also matters when the number of stimuli is reduced to five health profiles (C and D) rather than eight (A, B and E) and to analyse to what extent the elicitation of context free utilities is invariant to this respect. In other two contexts (F and G) quality of life was kept constant (frequency value of quality of life is the same in F and G) and life years were changed. We expect common health profiles with different frequency values to be valued differently across contexts.

Comparison of different contexts allows us to interpret the source of the context effects:

- **Comparison of A-B, C-D, and F-G**. Range Frequency Theory predicts that value attached to a specific health profiles is higher when the frequency value is higher. For example, (Q222, 20), (Q322, 20) and (Q332, 20) have a higher frequency value in B than in A. Also (Q222, 20) has a higher frequency value in D than in C. In addition, (Q333, 15), (Q333, 14) and (Q333, 13) have a higher frequency value in F than in G. On the other hand RFT entails that, for a given health profile, the contextual valuation is the same when the frequency value is the same. For example, (Q543, 20) has the same frequency value across contexts A, B and E, on the one hand, and across C and D, on the other hand; therefore valuation should not vary. Any differences in the valuation of the latter cannot be accounted for by RFT. Finally, AL makes the same prediction than RFT except for (Q543, 20) that would have a higher valuation in B and D than in A and C, respectively. The reason is that the adaptation level (average value of health profiles) is lower in A and in C, respectively. Following this criterion valuation in E should be higher than in A too.

- **Comparison of B-E**. In this comparison we can further test RFT with respect to AL. Two common health profiles differ in the frequency value. RFT predicts that the contextual value of (Q222, 20) is higher in B while the contextual value of (Q433, 20) is higher in E. However AL cannot make this type of predictions because it entails that the value should be higher in that context, either B or E, where the average value of contextual stimuli is lower.
- **Comparison of A-E**. In this case the four common health profiles have the same frequency value in both contexts. Therefore RFT predicts no differences in valuation. However AL would entail that the average value of contextual health profiles is lower in E and therefore valuation should be higher in that context. Specifically we can test whether valuation of (Q211, 20), (Q221, 20), (Q222, 20) and (Q543, 20) changes when they are valued jointly with (Q322, 20) and (Q332, 20), like in group A, rather jointly with (Q433, 20) and (Q443, 20), like in group E.

Table 2. HUI-3 levels of severity for ambulation, emotion and pain

Attribute	Description
Levels of ambulation	
1	Able to walk around the neighbourhood without difficulty, and without walking equipment.
2	Able to walk around the neighbourhood with difficulty; but does not require walking equipment or the help of another person.
3	Able to walk around the neighbourhood with walking equipment, but without the help of another person.
4	Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighbourhood.
5	Unable to walk alone, even with walking equipment. Able to walk short distances with the help of another person, and requires a wheelchair to get around the neighbourhood.
6	Cannot walk at all.
Levels of emotion	
1	Happy and interested in life.
2	Somewhat happy.
3	Somewhat unhappy.
4	Very unhappy.
5	So unhappy that life is not worthwhile.
Levels of pain	
1	Free of pain and discomfort.
2	Mild to moderate pain that prevents no activities.
3	Moderate pain that prevents a few activities.
4	Moderate to severe pain that prevents some activities.
5	Severe pain that prevents most activities

Note. In case of Full Health (Q111) in which level was 1 for the three attributes we described the health profiles as "*no health problems*".

Table 3. Health profiles and frequency values ($\boldsymbol{F_i}$) for each context

Quality of life manipulation						**Life years manipulation**		
			$\boldsymbol{F_i}$				$\boldsymbol{F_i}$	
Health profiles	**A**	**B**	**C**	**D**	**E**	**Health profiles**	**F**	**G**
(Q111, 20)	7/7	7/7	4/4	4/4	7/7	(Q111, 20)	7/7	7/7
(Q211, 20)	6/7		3/4		6/7	(Q333, 18)	6/7	
(Q221, 20)	5/7				5/7	(Q333, 17)	5/7	
(Q222, 20)	4/7	6/7	2/4	3/4	4/7	(Q333, 16)	4/7	
(Q322, 20)	3/7	5/7				(Q333, 15)	3/7	6/7
(Q332, 20)	2/7	4/7				(Q333, 14)	2/7	5/7
(Q333, 20)		3/7		2/4		(Q333, 13)	1/7	4/7
(Q433, 20)		2/7			3/7	(Q333, 12)		3/7
(Q443, 20)					2/7	(Q333, 11)		2/7
(Q543, 20)	1/7	1/7	1/4	1/4	1/7	(Q333, 10)		1/7
Death	0/7	0/7	0/4	0/4	0/7	Death	0/7	0/7

3.3. Evaluation Method: Time Trade Off

The TTO exercise for contexts A to E (quality of life manipulation) consisted in a ranking task in which respondents ordered the contextual health profiles to be valued (see Table 3) along with ten more health profiles consisting in (Q111, X) with X=18, 16, 14, 12, 10, 8, 6, 4, 2, 0. The ranking task is illustrated in Figures 1 and 2. In the first place individuals are shown a situation like in Figure 1 where the health profiles to be valued were represented by brown cards that included duration (*esperanza de vida*) and quality of life (*calidad de vida*). The health profiles are fixed and ordered by severity, the best at the top and the worst at the bottom. Then new health profiles like (Q111, X) appear at the bottom of the screen represented by blue cards. Respondents were instructed to locate this new (Q111, X) between the brown cards according to their preferences. For example, if a respondent considered (Q111, 18) to be worse than (Q211, 20) and better than (Q221, 20) it should be located in between them. The ranking exercise ended up with a situation like that represented in Figure 2. In any moment blue cards could be moved and rearranged by participants until they considered the order to be appropriate. Notice that we considered Death to be part of the contextual stimuli by including a blue card (described as *0 years, Immediate death*; in Spanish: "*0 años, Muerte inmediata*") rather than including a fixed brown card. This was the case for the context A to E (quality of life manipulation). The reason is because we wanted individual to be able to evaluate some health profiles to be worse than death. Nonetheless the TTO exercise in contexts F and G (life years manipulation) Death was represented by a brown fixed card at the bottom of the screen because it was considered that any health profile with quality of life Q333 would be preferred to immediate Death.[1]

The eventual goal of TTO is to achieve to that number of years in Full Health (Q111) that are equivalent to each health profile. We use the TTO ranking to extrapolate that value as the middle point between the lowest blue card considered better and the highest blue card considered worse. For example in Figure 2 (Q221, 20) is ordered in between (Q111, 18) and (Q111, 16). Therefore we consider $(Q221, 20) \sim (\text{Q111}, 17)$, this is 20 years in Q221 is equivalent to 17 years in Full Health. In the case of Health profiles that were located below immediate death (i.e. worse than death health profiles) it was given a value of -1.[2]

The TTO ranking exercise was preceded by another ranking exercise in which individuals ordered the brown cards (health profiles) according to severity. The objective was to make subjects to be familiar with the health profiles for valuation and therefore with the context. Also, information and explanations were available to respondents from the beginning of the online survey. For example they were explained that a health profile could vary in quality of life or in life expectancy. They also were shown the rationale of a TTO and some example of decisions in which they traded quality of life for life expectancy. They were requested to consider the hypothetical situations as if they were real and respond according to their preferences.

[1] In a pilot survey we made respondents to repeat the same TTO exercise one more time. However responses were not significantly different in the two repetitions. So in the main study only one single TTO exercise was included.

[2] Given that we were not interested in valuation of worse than death health profiles our method did not have a lower bound in those cases and then we assigned -1. However, only 4.74% of times a health profile was considered as worse than death by some of the respondents.

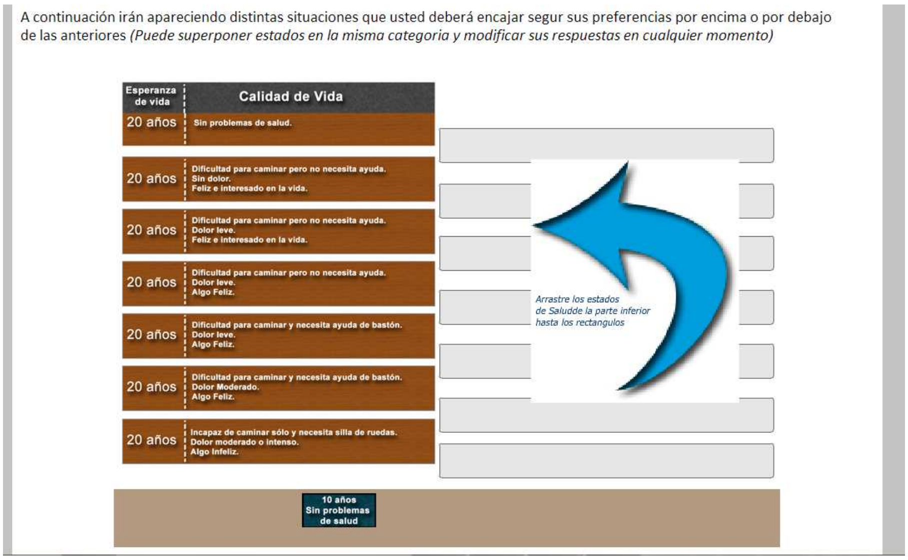


Figure 1. TTO exercise. Health profiles represented by fixed brown cards

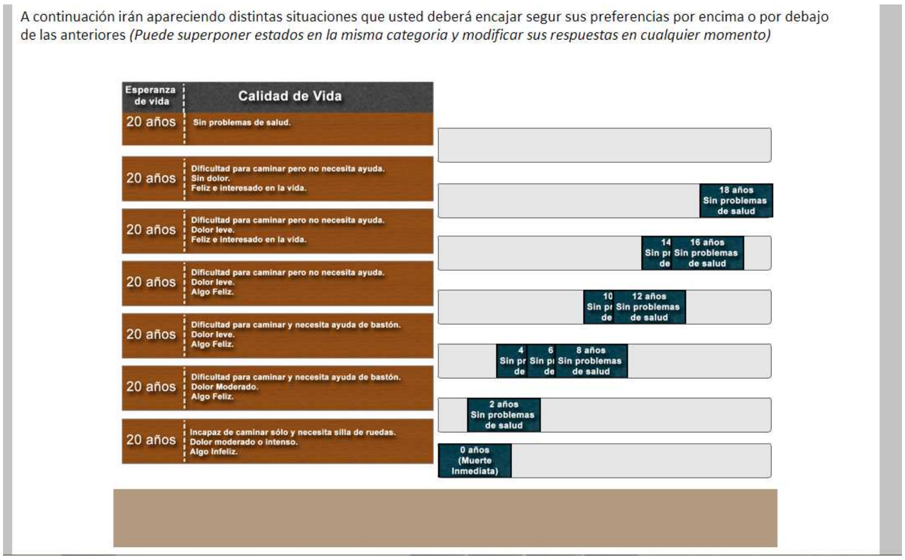


Figure 2. TTO exercise. Example of ordering

4. Results

For the ease of exposition we present results of manipulation of quality of life and life years separately.

4.1. Quality of life manipulation

Context-dependent utilities (expression 4) and context-free utilities (expression 7) are computed considering that life years are discounted by an exponential discount function, i.e. $D(T) = \frac{1-(1-r)^T}{r}$. Where r is the discount rate and T is the life expectancy for the health profile. We use $r = 0.05$ which is similar to previous elicitations (Attema et al., 2012) and to discount rates used by medical treatments evaluation agencies.

In addition for the estimation of context-free quality of life it is necessary to estimate the RDV weighting parameter. We obtain $\omega = 0.7$ by minimizing the sum of *MEAN* absolute deviations between the mean context-free utility for each quality of life and context and the mean context-free utilities for each quality of life, i.e. minimizing $\sum_i MEAN\left|Q_i^k - \overline{Q}_\iota\right|$ where Q_i^k is the mean context-free utility of quality of life i computed for context k and $\overline{Q}_\iota$ is the mean context-free utility of i for all contexts that include it.

Mean figures for context dependent utilities are in Table 4 and context free utilities are shown in Table 5. Statistical tests are performed to test whether utilities are equal across contexts. For that purpose we use the Kruskal-Wallis (K-W) non parametric equality of distribution test and Wilks' lambda (W' l) equality of mean test for multiple comparisons. We also perform pair wise Wilcoxon equality of distribution test and equality of mean test based on non parametric bootstrapping. We use both, equality of distribution and mean tests, because we found that for some comparisons utility distributions were statistically different while means do not differ significantly. We also compare graphically utilities in Figure 3 for contexts A vs B, C vs D, B vs E and A vs E. Results for these pair wise comparisons are:

- **Comparison of A-B.** Common qualities of life to both contexts are: Q222, Q322, Q332 and Q543. It can be seen in Figure 3 and Table 4 that context-dependent utilities for Q222, Q322 and Q332 were higher in Context B as predicted by RDV and AL. Utilities in A were 0.71, 0.64 and 0.55, respectively, while in context B were 0.79, 0.73 and 0.65, respectively. Pair wise Wilcoxon and bootstrap tests conclude that differences are statistically significantly different (*p-value*<0.05). On the contrary, context-free utilities are nearly identical and no statistical differences were found by bootstrap test (p-values are respectively 0.98, 0.99 and 0.76). Wicoxon *tests* cannot reject equality of distribution for Q222 and Q322 (p-values 0.63 and 0.16 respectively) while for Q332 it is rejected at 5% (notice the reduction of significativity even in this latter case). This result is consistent with the need to control for range-frequency effects to obtained true underlying preferences. Nonetheless, we found that utility for Q543 is higher in context A while no difference are predicted by RDV and AL predicts the opposite result. Although the difference is not significant at 5% of error this finding suggests that there may sources of context effects distinct to those considered in this study. Even more computation of context-free utilities according to RDV (expression 7) does not lower the gap.
- **Comparison of C-D.** In this case, there are two health states common to both contexts: Q222 and Q543. Consistently with RDV and AL utility for Q222 was significantly

higher in D than in C, 0.69 vs 0.60 (Wilcoxon test and bootstrap *tests p-value*<0.05). However differences disappear between mean context-free figures, 0.65 vs 0.66, and are not significant according to bootstrap mean test (*p-value*=0.77). In this case, Wicoxon test rejects equality of distribution of context-free utilities in disagreement with the mean test. This result is plausible because controlling for range frequency effects affect mainly to the centre of the distribution and not to the shape of it. In addition it suggests that context effects may be changing the shape of distribution of responses.[3] With respect to Q543 context-dependent utility is higher in C (0.39) than in D (0.37) but not statistically significant as predicted by RFT. This is not consistent with AL; this theory predicts higher context dependent utilities in D than in C for Q543.

- **Comparison of B-E.** Following differences in frequency values (Table 3) context dependent utility for Q222 was higher in B than in E (Wilcoxon and bootstrap *tests p-value*<0.01). Also consistently with range frequency theory Q433 was valued higher in E than in B (Wilcoxon and bootstrap *tests p-value*<0.01). This finding allows us to recognize RFT as a better account for contexts effects than AL. Once we control for differences in frequency value, context-free utilities do not differ for Q222 between contexts (Wilcoxon p-value=0.37 and bootstrap p-value=0.71). However we cannot explain differences in valuation of Q433 by computing context-free values (*p-value*<=0.05).[4]
- **Comparison of A-E.** Context-dependent utilities for Q211, Q221 and Q222 are the same in both contexts as predicted by RFT. However context dependent utility for Q543 is again higher in A than in E (not significant a 5% of error). As in the comparison of contexts A-B this result suggests that there may be some contextual variables that make Q543 to be perceived differently in context A. Nonetheless, if we analyse graphically utilities for A and E in Figure 3 we can observe that context free utilities follow a more rational pattern. For example, context dependent utility for Q433 is higher than for Q332 while the former is clearly a worse quality of life. In case of context-free utilities the ordering is the other way around, as rationality suggests.

With the previous comparisons we can acknowledge that the computation of context-free utilities by applying RDV model in expression 7 makes valuations to be more consistent between contexts. However another different issue is whether these “context-free” values are really context-invariant. The results described above suggest that the answer is no given that context free utilities vary between contexts for Q433 (B vs E) and Q543 (A vs B, A vs E). However, those results are based on pair wise comparisons between contexts with the same number of stimuli (A, B and E on the one hand, and C and D on the other hand). To further address this question we can observe Kruskal-Wallis equality of distribution test and Wilks' lambda mean test for multiple comparisons of all contexts together for each health state. Both tests conclude that context free utilities are significantly different also for Q211 and Q222 (p-value<0.05). However, no significant differences were found for Q543 so that we can conclude that significance of pair wise comparison tests are due to random errors. Specifically, context free utilities for Q211 and Q222 were significantly higher in contexts where the number of

[3] More detailed analysis of context-free utilities tells us that the distribution in context D is more negatively skewed than in C (skewness coefficient -1.65 vs -0.94) more dispersed (variance 0.13 vs 0.11) and has a higher frequency of responses in the tails (kurtosis coefficient 5.84 vs 5.18).

[4] When ω take a value of 0.18, differences between B and E for Q433 disappear. However context-free utility of Q433 turns to be 1.25; an implausible value higher than the utility of Full Health.

stimuli is high, eight (A, B and E) rather than five (C and D) with p-value<=0.1 (see pair wise comparison Wilcoxon and bootstrap tests in Table 5).[5]

To further check whether the results are driven by the use of RDV instead of applying range frequency operating at the response level we also computed context-free utilities using R-RFT in expression 8. Results are shown in Table A1 in appendix. Context-free utilities are consistently lower than those computed by RDV for each health state and context (0.06 utility points on average). However, context free utilities were significantly different across context for Q211, Q222, and Q433 with p-value<0.1 according to Kruskal-Wallis and Wilks' lambda tests. Again, context free utilities for Q211 and Q222 were significantly higher in contexts A, B and E than in C and D (pair wise comparison Wilcoxon and bootstrap test significant with p-value<0.1).

[5] Variation of ω did not make context-free utilities to match for Q211 and Q222. Even when considering a different weighting parameter for contexts with eight, on the one hand, and five health profiles, on the other hand.

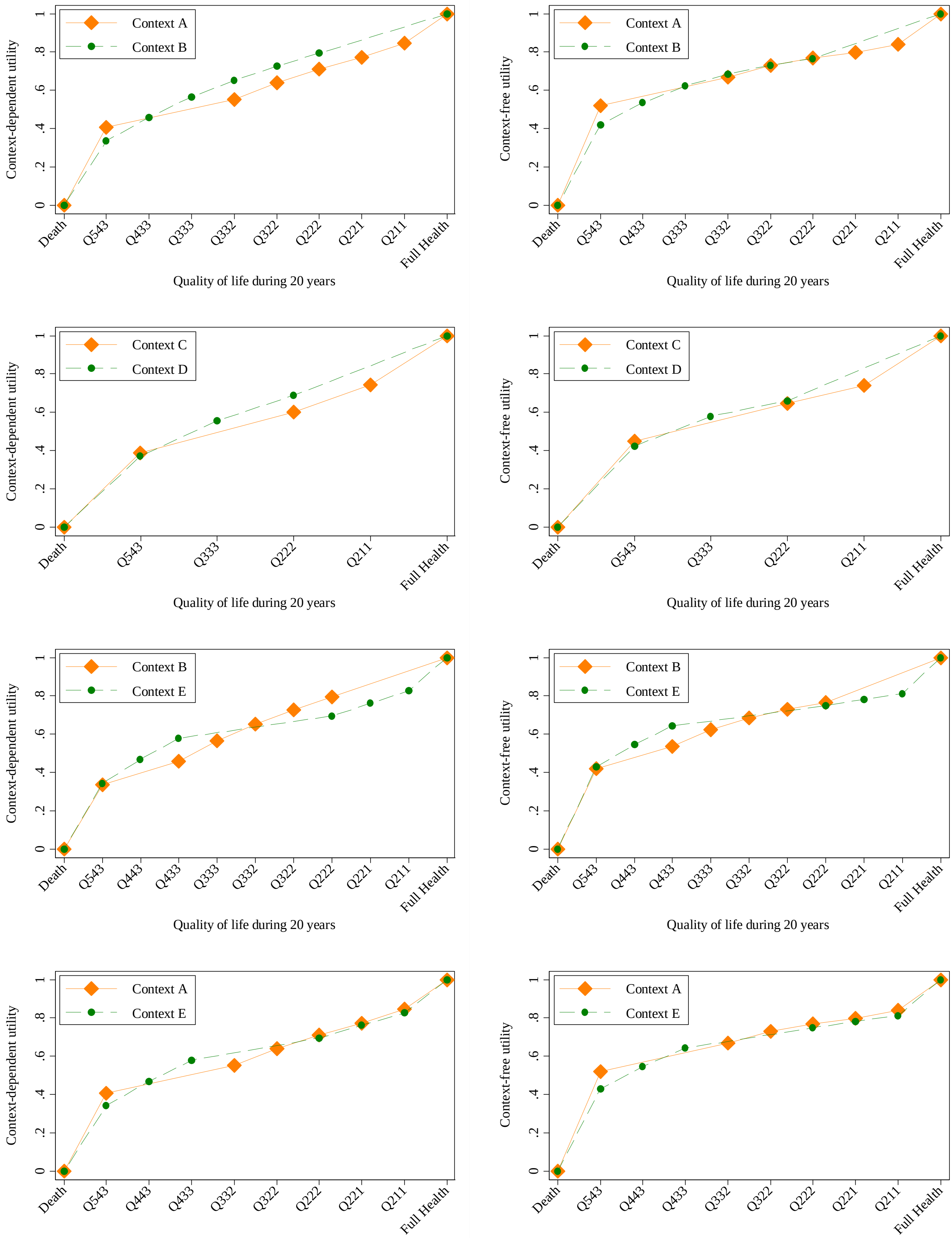


Figure 3. Mean context-dependent (left) and context-free (right) utilities by context. Quality of life manipulation. ($\omega = 0.7$)

Table 4. Mean context dependent utilities by context and statistical tests. 5% discount rate

	Utilities ($Q_{i,k}$)						
Quality of life for 20 years	**A**	**B**	**C**	**D**	**E**	**K-W (*p-value*)**	**W' l (*p-value*)**
Q211	0.85		0.74		0.83	0.00	0.00
Q221	0.77				0.76	0.71	0.76
Q222	0.71	0.79	0.60	0.69	0.70	0.00	0.00
Q322	0.64	0.73				0.00	0.02
Q332	0.55	0.65				0.00	0.01
Q333		0.57		0.56		0.99	0.79
Q433		0.46			0.58	0.00	0.00
Q443					0.47		
Q543	0.41	0.34	0.39	0.37	0.34	0.03	0.27

Quality of life for 20 years	**Wilcoxon ranksum tests (*p-value*)**									
	A vs B	**C vs D**	**B vs E**	**A vs E**	**A vs C**	**A vs D**	**B vs C**	**B vs D**	**C vs E**	**D vs E**
Q211				0.74	0.00				0.00	
Q221				0.71						
Q222	0.00	0.00	0.00	0.79	0.00	0.79	0.00	0.00	0.00	0.97
Q322	0.00									
Q332	0.00									
Q333								0.99		
Q433			0.00							
Q543	0.07	0.74	0.88	0.09	0.42	0.53	0.03	0.03	0.029	0.04

Quality of life for 20 years	**Non-parametric bootstrap equality-of-mean tests (*p-value*)**									
	A vs B	**C vs D**	**B vs E**	**A vs E**	**A vs C**	**A vs D**	**B vs C**	**B vs D**	**C vs E**	**D vs E**
Q211				0.48	0.00				0.00	
Q221				0.75						
Q222	0.01	0.01	0.00	0.69	0.00	0.52	0.00	0.00	0.00	0.81
Q322	0.02									
Q332	0.01									
Q333								0.79		
Q433			0.00							
Q543	0.06	0.61	0.86	0.08	0.61	0.36	0.15	0.37	0.19	0.44

Table 5. Mean context free utilities by context and statistical tests. 5% discount rate. RDV ($\boldsymbol{\omega = 0.7}$)

	Utilities (Q_i)						
Quality of life for 20 years	**A**	**B**	**C**	**D**	**E**	**K-W (*p-value*)**	**W' l (*p-value*)**
Q211	0.84		0.74		0.81	0.02	0.04
Q221	0.80				0.78	0.72	0.76
Q222	0.77	0.77	0.65	0.66	0.75	0.00	0.02
Q322	0.73	0.73				0.16	0.98
Q332	0.67	0.69				0.04	0.77
Q333		0.62		0.58		0.01	0.41
Q433		0.53			0.64	0.00	0.05
Q443					0.55		
Q543	0.52	0.42	0.45	0.42	0.43	0.38	0.28

Quality of life for 20 years	**Wilcoxon ranksum tests (*p-value*)**									
	A vs B	**C vs D**	**B vs E**	**A vs E**	**A vs C**	**A vs D**	**B vs C**	**B vs D**	**C vs E**	**D vs E**
Q211				0.74	0.01				0.02	
Q221				0.71						
Q222	0.63	0.00	0.37	0.79	0.00	0.00	0.00	0.00	0.00	0.01
Q322	0.16									
Q332	0.04									
Q333								0.01		
Q433			0.00							
Q543	0.07	0.74	0.88	0.09	0.30	0.41	0.34	0.38	0.45	0.43

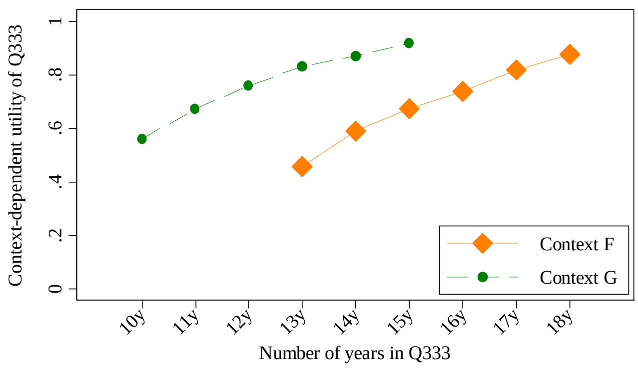
Quality of life for 20 years	**Non-parametric bootstrap equality-of-mean tests (*p-value*)**									
	A vs B	**C vs D**	**B vs E**	**A vs E**	**A vs C**	**A vs D**	**B vs C**	**B vs D**	**C vs E**	**D vs E**
Q211				0.48	0.01				0.10	
Q221				0.75						
Q222	0.98	0.77	0.71	0.68	0.01	0.02	0.01	0.03	0.04	0.07
Q322	0.99									
Q332	0.76									
Q333								0.39		
Q433			0.05							
Q543	0.07	0.61	0.85	0.09	0.17	0.06	0.56	0.97	0.71	0.88

4.2. Life years manipulation

In context F and G all health profiles are represented by a fixed quality of life (Q333) with different duration. If no context effects apply then all utilities should be the same. However the rest of the exposition shows that this is not the case. Results for context dependent and context free utility for that health state are shown in Figure 4 and Table 6. Again we assume 5% of discount rate. For estimation of R-RFT model weighting parameter we minimize the sum of absolute differences between context free utilities computed from responses to the three common health profiles in F and G: (Q333, 15), (Q333, 14) and (Q333, 13). Optimization resulted in $\omega = 0.51$.

As can be observed in Figure 4 context-dependent utilities derived from valuation of (Q333, 15), (Q333, 14) and (Q333, 13) are lower in context F as predicted by RFT. Differences are statistically significant (Wilcoxon and bootstrap tests p-value<0.01, see Table 6). This result is explained because the frequency value is lower in that group. Also, context dependent utilities of Q333 vary within contexts. In particular utilities computed from valuation of health profiles with a short duration (e.g. 10 years with Q333 in context G) are lower than utilities elicited from health profiles with long duration (e.g. 15 years with Q333 in context G). R-RFT in expression 6 can predict differences of this type since the former health profiles have a low rank ($Rank_{i,k}$) and therefore responses ($T_{FH,k}$) are low and eventually context-dependent utility of health states can have a low value.

On the other hand, computed context free utilities are much more similar to each other as can be observed in Figure 4. Context free utilities derived from (Q333, 15), (Q333, 14) and (Q333, 13) are not significantly different between F and G (Wilcoxon and Bootstrap p-value>=0.07). Even more, we test whether mean context free utilities vary across the 12 groups ($2 \text{contexts} \times 6 \text{health profiles}$) considered and no significant differences are found (Wilks' lambda p-value=0.99). In the meantime this test rejects equality of means of context dependent utilities (p-value<0.01). An illustrative fact is that standard deviation of mean values for context free utilities for the 12 groups is 33.2% of standard deviation of mean values for context dependent utilities. Therefore controlling for range frequency effect considerably reduces variation in elicited utilities and therefore inconsistency between contexts. However if we look at equality of distribution test (K-W) we reject the equality of context-free utilities for the 12 groups considered (p-value<0.006). This makes evident that RFT do not work very good when we have also to control for changes in the shape of distribution of responses.



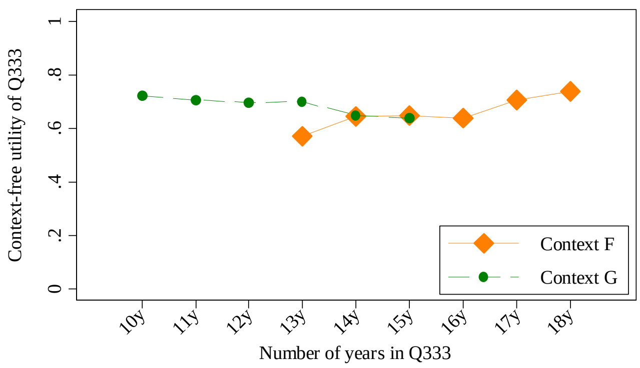


Figure 4. Mean context-dependent (left) and context-free (right) utilities by context. Life years manipulation. $\boldsymbol{\omega = 0.51}$

Table 6. Mean context dependent and context free utilities by context and statistical tests. 5% discount rate. R-RFT

Context-dependent utility of Q333				
			Tests	
Number of years Q333	**in F**	**G**	**Wilcoxon**	**Bootstrap.**
10y		0.56		
11y		0.67		
12y		0.76		
13y	0.46	0.83	0.00	0.00
14y	0.59	0.87	0.00	0.00
15y	0.68	0.92	0.00	0.00
16y	0.74			
17y	0.82			
18y	0.88			
			K-W	0.000
Stand. Deviation	*0.137*		**W' L**	0.000

Context-free utility of Q333 ($\omega = 0.51$)				
			Tests	
Number of years Q333	**in F**	**G**	**Wilcoxon**	**Bootstrap.**
10y		0.72		
11y		0.71		
12y		0.70		
13y	0.57	0.70	0.07	0.29
14y	0.65	0.65	0.42	0.98
15y	0.65	0.64	0.19	0.94
16y	0.64			
17y	0.71			
18y	0.74			
			K-W	0.006
Stand. Deviation	*0.0455*		**W' L**	0.99

5. Discussion

Range frequency theory turns to be a good account for shifts in valuation between contexts with the same number of health profiles. It predicts properly that valuation of a health profile relate positively to its frequency value. This is illustrated in the results presented graphically in Figures 3 and 4. Also if we regress TTO responses on frequency values including binary variables for each health profile (i.e. we estimate $T_{FH,k} = \alpha + \beta F_i + HP_i + \varepsilon$) we find a highly significant effect of frequency value (p-value<0.001). Indeed we find the same kind of effect when the frequency value of a health profile was varied either by quality of life or life years manipulation. This is, for a given health profile the number of years in Full Health equivalents are higher when the relative rank is higher. As a consequence elicited utilities for health states associated to health profiles are higher.

Results are consistent to repetitions of valuations by respondents. In a pilot study respondents repeated the TTO exercise in the same online survey in order to test to what extent contexts effects may be changed. In this pre-test study we restrict the analysis to 40 subjects for each of the next contexts: A, B, E, F and G. Comparison of first and second responses are shown in Table A2, for manipulation of quality of life, and A3, for manipulation of life years, in Appendix. As it can be observed in both cases utilities derived from responses follow the same pattern; i.e. utility of a health state is higher in those contexts where the frequency value is higher. Indeed, mean utilities barely vary between first and second valuation (no statistical significance; Wilcoxon test). The percentage of individuals giving the same response to both TTO exercises for a specific health profile and context is between 53 and 69%, for the quality of life manipulation case (contexts A, B and E), and between 51 and 81%, for the life years manipulation case (contexts F and G). Table A2 and A3 give us an approximation to the importance of context effects relative to regular errors committed by respondents even within the same context. For example, the average within context deviation between utilities derived from first and second responses for Q222, Q322 and Q332 in group A and B (0.013) is much lower than average context effects between contexts A and B (0.09).

Moreover RFT is a better explanation than other contrast effect theories like Adaptation Level Theory. Changes in the direction of valuation shifts like resulted in the comparison of groups B and E for Q222 and Q433 cannot be explained by differences in a single reference point between contexts as proposed by AL. In this respect valuation in the health domain do not differ to previous findings in psychophysical contextual judgment (Parducci, 1965; Birnbaum, 1974 and 1999; Parducci and Wedell, 1986; Wedell, 2008).

It is well known that Health States values vary across methods or elicitation procedures. For example, TTO lead to different utilities to that derived from Standard Gamble (Stiggelbout et al., 1994; Beichrodt, 2002) and from rating scales (Bleichrodt and Johannesson, 1997b). More closely in Chapter 1 of this thesis we analyse inconsistencies between Standard Gamble and Contingent valuation of road injuries. In Chapter 2 it was shown that evaluation mode, either Separate of Joint, affect valuation as well. Results in the present paper posit a new dilemma for inconsistency of health valuations. Not only evaluation procedures affect valuation but also the distribution of health states included in the evaluation context.

At least two possible reactions to this fact exist. One may consider that there is a true utility for each health state that is contaminated by the context and the main task of the analyst is to create procedures to eliminate biases for elicitation of the true underlying value. On the other hand, one may acknowledge that there is no such context free utility and the task of the analyst is to create the proper context for valuation. In this study we analyse the former possibility. We computed context free utilities based on RFT that turned to be more consistent between contexts with the same number of stimuli (see Figure 3 and Figure 4). However utility inconsistencies between contexts with a different number of contextual health states were not eliminated by the computation of context free utilities. Any between-groups differences in context free utilities cannot be explained by RFT. It is usual to change the weighting parameter to fit data (Parducci and Wedell, 1983) however this is not the case here (see footnote 5 above).

When there is no possibility for correcting contextual biases the latter reaction is a more plausible way out to contextual effects. In this case the main challenge is to normatively define the characteristics of an appropriate context for valuation. For example, a sensible criterion would be to create the valuation context as similar as possible to the real context in which health improvements (or losses) given by public policy are going to be experienced.

Two further comments can be said. First, it would be interesting to investigate to what extent the results encountered here for context effects also appear for different TTO procedures. Specifically, we used a ranking exercise to elicit TTO utilities, however we do not investigate whether TTO provided as choices between alternatives could improve consistency of valuation across contexts. A more simple variation of the procedure is to allow respondents to answer more accurate responses (i.e. life years in Full Health indifferent to the health profile to be valued). In our case we infer utilities from the ranking TTO responses to some good approximation but still these are mainly categorical or interval utilities. While for elicitation of average utilities this does not seem to be problematic it could be when we want to compare distribution of utilities.[6] This could explain why sometimes we can make average context-free utilities from different contexts equal while the shape of the distribution seems to not be the same. In Figure A1 in appendix it is shown distribution of context dependent (left) and context-free (right) utilities for some health states and comparing between contexts. It can be observed that distribution of bias-corrected utilities is more similar between contexts than distribution of context-dependent valuations. However, even in that case the distribution seems to be different. Wilcoxon test rejects equality of distribution at 5% only for Q332, between context A and B, and for Q222, between context C and D (see pair wise tests for context-free utilities in Table 5).

Eventually, in future studies it could be considered the idea that the contextual stimuli are not only a result of the experimental manipulation but a mix between this and the stimuli that respondents are used to encounter in their daily life. For example, the final context of valuation would be different for a person that works in a hospital than for a respondent working in an administrative office. The study of how eventual contextual stimuli are formed could allow us to make better predictions about experimental manipulation and bias-correction.

---

[6] For example, imagine an extreme case where only two utility intervals are possible with 0.5 and 0.7 as respective midpoints. Suppose also that in context A distribution of responses are such that all context dependent utilities are within the first interval, so 100% of responses are 0.5. Suppose now that in context B the distribution of context dependent utilities shifts to the right such that 50% of the sample continues within the first interval (utility 0.5) and 50% changes to the second (utility 0.7). If we correct utilities for average bias (0.1) by lowering responses in context B, context-free utilities would be 0.4 and 0.6 (for each half of the sample) not coinciding with any of the utilities in context A (100% are 0.5).

6. Conclusion

We manipulated contexts for evaluation of health profiles by changing the distribution of quality of life and life years associated to each health profile. Contexts had either eight or five health profiles. Basically, frequency values of health profiles were different across contexts. Significant higher values were found for health profiles with higher frequency values consistently with range frequency theory. Application of this model works properly to eliminate biases between contexts with the same number of health. However, even the computation of context free utilities did not eliminated inconsistencies between contexts with different number of health profiles. Context effects due to the distribution of contextual stimuli are a challenge for health utilities elicitation demanding new procedures for estimation of underlying true utilities or demanding the creation of appropriate contexts for valuation.

Appendix

Context free utilities under RDV.

Substituting expression (4) in (5) we have

$$\frac{D(T_{FH,k})}{D(T_i)} = \omega \times \frac{Q_i - Q_{min,k}}{Q_{max,k} - Q_{min,k}} + (1-\omega) \times \frac{Rank_{i,k} - 1}{N_k - 1}. \qquad \text{(A1)}$$

Given $F_{i,k} = \frac{Rank_{i,k}-1}{N_k-1}$, $Q_{min,k} = 0$ and $Q_{max,k} = 1$ and rearranging A1 we have:

$$\frac{D(T_{FH,k})}{D(T_i)} - (1-\omega)F_{i,k} = \omega \times Q_i;$$

$$Q_i = \frac{D(T_{FH,k})}{D(T_i)\omega} - \frac{(1-\omega)}{\omega} F_{i,k}. \qquad \text{(7)}$$

Context free utilities under RFT.

Given $F_{i,k} = \frac{Rank_{i,k}-1}{N_k-1}$, $T_{min,k} = 0$ and $T_{max,k} = 20$ expression (6) converts into:

$$T_{FH,k} = \left[\omega \times \frac{T_{FH}}{20} + (1-\omega)F_{i,k}\right] 20. \qquad \text{(A2)}$$

And rearranging,

$$T_{FH} = \left[\frac{T_{FH,k}}{20\omega} - \frac{(1-\omega)}{\omega} F_{i,k}\right] 20. \qquad \text{(A3)}$$

Substituting expression (A3) in (3) we have

$$Q_i = \frac{D\left(\left[\frac{T_{FH,k}}{20\omega} - \frac{(1-\omega)}{\omega} F_{i,k}\right] 20\right)}{D(T_i)}. \qquad \text{(8)}$$

Table A1. Mean context free utilities by context and statistical tests. 5% discount rate. R-RFT ($\omega = 0.68$)

Quality of life for 20 years	**Utilities (Q_i)** A	**B**	**C**	**D**	**E**	**K-W (*p-value*)**	**W' l (*p-value*)**
Q211	0.80		0.68		0.76	0.02	0.05
Q221	0.73				0.71	0.71	0.76
Q222	0.70	0.70	0.57	0.58	0.67	0.00	0.053
Q322	0.66	0.66				0.06	0.99
Q332	0.60	0.60				0.01	0.94
Q333		0.54		0.50		0.01	0.43
Q433		0.47			0.56	0.01	0.084
Q443					0.48		
Q543	0.47	0.37	0.39	0.36	0.38	0.38	0.24

Quality of life for 20 years	**Wilcoxon ranksum tests (*p-value*)** A vs B	C vs D	B vs E	A vs E	A vs C	A vs D	B vs C	B vs D	C vs E	D vs E
Q211				0.74	0.02				0.02	
Q221				0.71						
Q222	0.05	0.00	0.048	0.79	0.00	0.008	0.00	0.00	0.008	0.02
Q322	0.06									
Q332	0.01									
Q333								0.01		
Q433			0.009							
Q543	0.07	0.74	0.88	0.098	0.29	0.41	0.34	0.38	0.45	0.44

Quality of life for 20 years	**Non-parametric bootstrap equality-of-mean tests (*p-value*)** A vs B	C vs D	B vs E	A vs E	A vs C	A vs D	B vs C	B vs D	C vs E	D vs E
Q211				0.44	0.009				0.13	
Q221				0.76						
Q222	0.90	0.89	0.62	0.69	0.02	0.05	0.03	0.06	0.07	0.13
Q322	0.99									
Q332	0.94									
Q333								0.43		
Q433			0.07							
Q543	0.07	0.56	0.83	0.097	0.13	0.04	0.72	0.80	0.89	0.65

Table A2. Pilot study of quality of life manipulation. Individuals repeat TTO responses

Quality of life for 20 years	Utilities (first responses)			Utilities (second responses)			Number of subjects with the same responses (%)		
	Contexts			Contexts			Contexts		
	A	B	E	A	B	E	A	B	E
Q211	0.86		0.80	0.87		0.83	63		69
Q221	0.80		0.74	0.80		0.76	63		59
Q222	0.75	0.84	0.68	0.74	0.83	0.66	53	68	59
Q322	0.67	0.76		0.68	0.75		55	53	
Q332	0.58	0.70		0.58	0.66		60	60	
Q333		0.63			0.59			53	
Q433		0.55	0.55		0.50	0.55		55	56
Q443			0.45			0.46			56
Q543	0.42	0.44	0.35	0.40	0.37	0.32	65	55	67

Note 1. In the pilot study 40 subjects were assigned to each context.
Note 2. Contexts C and D were not analysed in the pilot study.

Table A3. Pilot study of life years manipulation. Individuals repeat TTO responses

Number of years in Q333	Utilities (first responses)		Utilities (second responses)		Number of subjects with the same responses (%)	
	Contexts		Contexts		Contexts	
	F	G	F	G	F	G
10y		0.56		0.53		81
11y		0.67		0.65		81
12y		0.76		0.74		71
13y	0.46	0.83	0.45	0.82	61	74
14y	0.59	0.87	0.58	0.86	59	71
15y	0.68	0.92	0.67	0.91	56	76
16y	0.74		0.74		54	
17y	0.82		0.83		54	
18y	0.88		0.88		51	

Note 1. In the pilot study 40 subjects were assigned to each context.

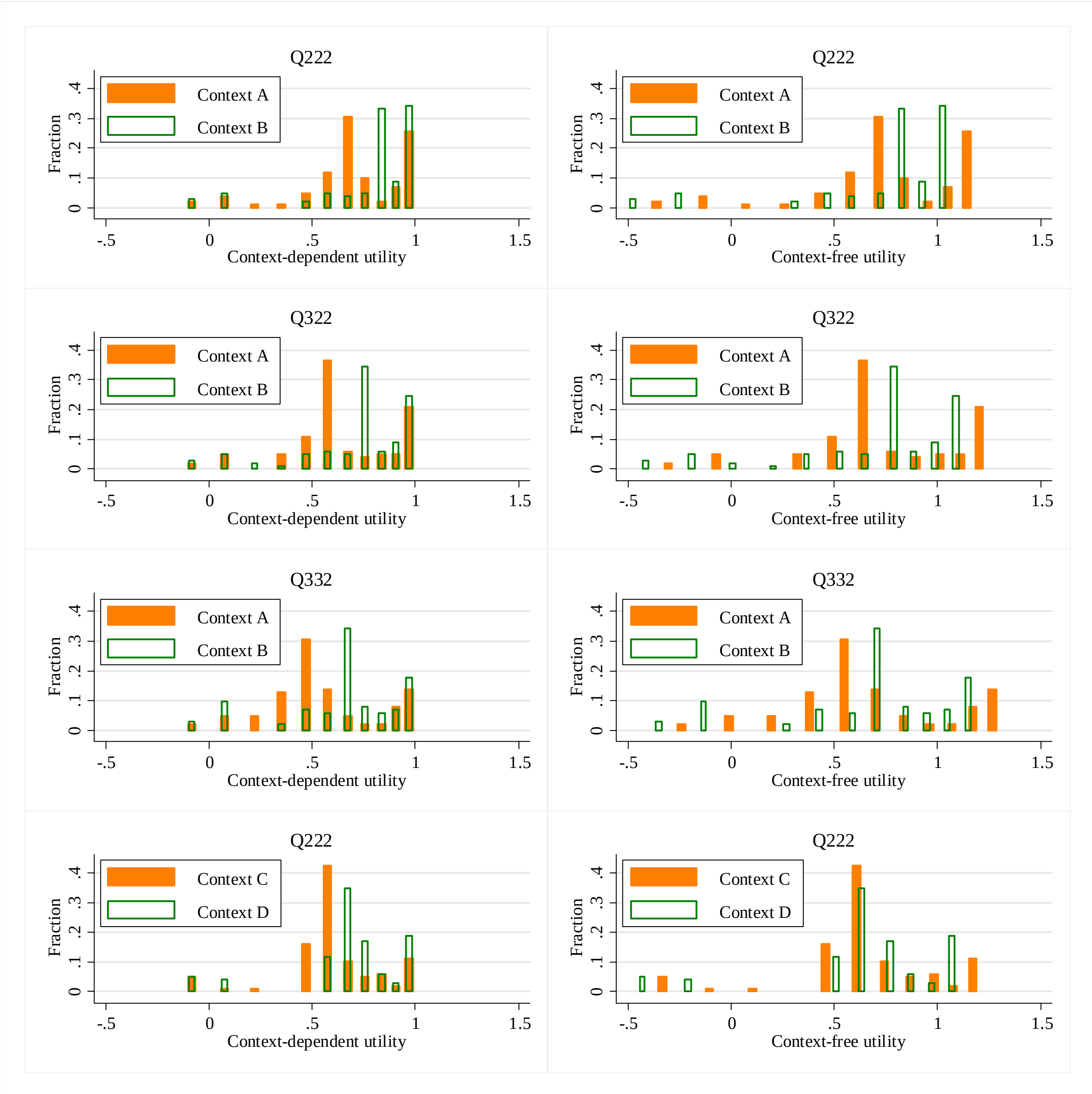


Figure A1. Distribution of context-dependent (left) and context-free (right) utilities by context.

References

1. Attema, A. E., Bleichrodt, H., and Wakker, P. P. (2012). A direct method for measuring discounting and QALYs more easily and reliably. *Medical Decision Making*, *32*(4), 583-593.
2. Birnbaum, M. H. (1999). How to show that 9> 221: Collect judgments in a between-subjects design. *Psychological Methods*, *4*(3), 243.
3. Birnbaum, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception and Psychophysics*, *15*, 89–96.
4. Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health economics*, *11*(5), 447-456.
5. Bleichrodt, H., and Johannesson, M. (1997a). An experimental test of a theoretical foundation for rating-scale valuations. *Medical Decision Making*, *17*(2), 208-216.
6. Bleichrodt, H., and Johannesson, M. (1997b). Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. *Journal of Health Economics*, *16*(2), 155-175.
7. Brown, G. D., Gardner, J., Oswald, A. J., and Qian, J. (2008). Does Wage Rank Affect Employees' Well-being? *Industrial Relations: A Journal of Economy and Society*,*47*(3), 355-389.
8. Feeny, D., Furlong, W., Boyle, M., and Torrance, G. W. (1995). Multi-attribute health status classification systems. Health Utilities Index. *Pharmacoeconomics*, 7, 490-502.
9. Helson, H. (1964). Current trends and issues in adaptation-level theory. *American Psychologist*, *19*(1), 26.
10. Parducci, A. (1983). Category ratings and the relational character of judgment. *Advances in psychology*, *11*, 262-282.
11. Parducci, A. (1965). Category judgment: a range-frequency model. *Psychological review*, *72*(6), 407.
12. Parducci, A., and Wedell, D. H. (1986). The category effect with rating scales: number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(4), 496.
13. Pinto-Prades, J. L. (2013) Imprecise preferences and order effects in the time trade-off. Health Economics Study Group January 2013.
14. Rashotte, M. A., and Wedell, D. H. (2012). Context effects on tempo and pleasantness judgments for Beatles songs. *Attention, Perception, and Psychophysics*, *74*(3), 575-599.
15. Robinson, A., Loomes, G., and Jones-Lee, M. (2001). Visual analog scales, standard gambles, and relative risk aversion. *Medical Decision Making*, *21*(1), 17-27.
16. Ruiz, M., Rejas, J., Pardo, A., Rebollo, I. and Soto, J. (2003). Adaptación y validación del Health Utilities Index Mark 3 al castellano y baremos de corrección en la población española. *Medicina clínica*, *120*(3), 89-96.
17. Schwartz, A. (1998). Rating scales in context. *Medical Decision Making*, *18*(2), 236-236.
18. Simonson, I., and Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research (JMR)*, *29*(3).
19. Stiggelbout, A. M., Kiebert, G. M., Kievit, J., Leer, J. W. H., Stoter, G., and De Haes, J. C. J. M. (1994). Utility Assessment in Cancer Patients Adjustment of Time Tradeoff

Scores for the Utility of Life Years and Comparison with Standard Gamble Scores.*Medical Decision Making*, *14*(1), 82-90.

20. Torrance, G. W. (2005). *Methods for the economic evaluation of health care programmes*. M. F. Drummond (Ed.). Oxford university press.
21. Tversky, A., and Simonson, I. (1993). Context-dependent preferences. *Management science*, *39*(10), 1179-1189.
22. Watkinson, P., Wood, A. M., Lloyd, D. M., and Brown, G. D. (2013). Pain ratings reflect cognitive context: A range frequency model of pain perception. *Pain*, *154*(5), 743-749.
23. Wedell, D. H. (2008). A similarity-based range—frequency model for two-category rating data. *Psychonomic bulletin and review*, *15*(3), 638-643.
24. Wedell, D. H. (1995). Contrast effects in paired comparisons: Evidence for both stimulus-based and response-based processes. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(5), 1158.
25. Wedell, D. H., and Parducci, A. (1988). The category effect in social judgment: experimental ratings of happiness. *Journal of Personality and Social Psychology*, *55*(3), 341.
26. Wedell, D. H., Parducci, A., and Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *American Journal of Psychology*, *102*(233-248).
27. Wedell, D. H., Santoyo, E. M., and Pettibone, J. C. (2005). The thick and the thin of it: Contextual effects in body perception. *Basic and Applied Social Psychology*, *27*(3), 213-228.